



GRADO EN CIENCIA DE DATOS



VNIVERSITAT
DE VALÈNCIA

TRABAJO FIN DE GRADO

DETECCIÓN DE NOTICIAS FALSAS USANDO
PROCESADO DE LENGUAJE NATURAL

AUTOR: MIGUEL RICARDO CLARAMUNT ARGOTE

TUTOR: JOAN VILA FRANCÉS

JULIO 2023

Declaración de autoría:

Yo, Miguel Ricardo Claramunt Argote, declaro la autoría del Trabajo Fin de Grado titulado “Detección de noticias falsas usando Procesado de Lenguaje Natural” y que el citado trabajo no infringe las leyes en vigor sobre propiedad intelectual. El material no original que figura en este trabajo ha sido atribuido a sus legítimos autores.

Valencia, 11 de julio de 2023

Fdo: Miguel Ricardo Claramunt Argote

Resumen:

Las *fake news* son un fenómeno que ha tomado un gran impacto social estos últimos años y sus efectos tampoco son nimios en la población. Para reducir estos efectos, los medios de comunicación han promovido el *fact checking*, aunque este proceso es extremadamente costoso en todos los ámbitos. Por ello, sería de ayuda un método automático para agilizar este proceso.

Primeramente, definimos el concepto de *fake news* para tipificar el problema, objetivos y recursos para abordarlo.

En este trabajo usamos 3 *datasets* y 13 modelos de lenguaje basados en *transformers* para implementar una serie de clasificadores binarios. También aplicamos técnicas de *Explanable AI* (XAI), para comprobar si el entrenamiento sigue el progreso esperado.

Los mejores resultados se categorizan según diferentes métricas. En líneas generales los modelos LARGE suelen funcionar mejor que los BASE. También, funcionan mejor para detectar noticias verdaderas que falsas. Aplicando XAI concluimos que el aprendizaje no es el idóneo, funcionando solamente como esperábamos en 1 *dataset*.

Finalmente, se han analizado diferentes aspectos: interpretabilidad de modelos, limitaciones de estos y posibles sesgos en la clasificación.

Abstract:

Fake news are a phenomenon that gained significant impact on society in recent years – and its effects on the population are not negligible either. In order to reduce these repercussions, media outlets have advocated fact checking, although this process is extremely costly. Therefore, an automatic method to streamline this process would be beneficial.

First, we define the concept of fake news to typify the problem, objectives and resources to address it.

In this work we use 3 datasets and 13 language models based on transformers to implement a series of binary classifiers. We also apply Explanable AI (XAI) techniques to check if the training follows expected progress.

The best results are categorized according to different metrics. Generally speaking, LARGE models tend to perform better than BASE models. Also, they perform better for detecting true news than fake news. Applying XAI we conclude that learning is not ideal, performing only as expected in 1 dataset.

Finally, different aspects have been analyzed: model interpretability, limitations of the models and possible biases in the classification.

Resum:

Les *fake news* són un fenomen que ha pres un gran impacte social aquests darrers anys i els seus efectes tampoc són nimis a la població. Per reduir aquests efectes, els mitjans de comunicació han promogut el *fact checking*, encara que aquest procés és extremadament costós en tots els àmbits. Per això, seria ajuda un mètode automàtic per agilitzar aquest procés.

Primerament, definim el concepte de *fake news* per tipificar el problema, objectius i recursos per abordar-lo.

En aquest treball fem servir 3 *datasets* i 13 models de llenguatge basats en *transformers* per implementar una sèrie de classificadors binaris. També apliquem tècniques de *Explainable AI* (XAI), per comprovar si l'entrenament segueix el progrés esperat.

Els millors resultats es categoritzen segons diferents mètriques. En línies generals, els models LARGE solen funcionar millor que els BASE. També funcionen millor per detectar notícies veritables que falses. Aplicant XAI concloem que l'aprenentatge no és l'idoni, funcionant només com esperàvem a 1 *dataset*.

Finalment, s'han analitzat els diferents aspectes: interpretabilitat de models, limitacions d'aquests i possibles biaixos a la classificació.

Agradecimientos:

A mi familia, mi madre María Eugenia y mi padre José Ricardo, gracias por confiar en mi y apoyarme incondicionalmente.

A todas las personas que han estado a mi lado tanto en València como en Barcelona, gracias por formar parte y haber podido vivir estos años juntos.

Índice general

1. Introducción	15
1.1. Motivación	16
1.2. Objetivos	17
1.3. Organización de la memoria	18
2. Definiciones	19
3. Estado del arte	23
3.1. Antecedentes	23
3.2. Tecnologías	24
4. Análisis del problema	27
4.1. Estudio preliminar	27
4.2. Datasets	28
4.2.1. Descripción del conjunto de datos	28
4.2.2. Procesamiento del conjunto de datos	29
4.3. Modelos y clasificadores utilizados	31
4.3.1. Bag of Words	31
4.3.2. TF-IDF	31
4.3.3. Regresión logística	32
4.3.4. Naïve Bayes	32
4.3.5. SVM y SGD	33
4.3.6. Random Forest	33
4.3.7. BERT	33
4.3.8. DistilBERT	34
4.3.9. RoBERTa	34
4.3.10. DeBERTa	35
4.4. Material, recursos utilizados e implementación	35
5. Metodología aplicada	37

5.1. Entrenamiento de modelos estadísticos	37
5.2. Entrenamiento de transformers	37
6. Resultados obtenidos y evaluación	39
6.1. Resultados obtenidos	39
6.2. Interpretabilidad de los modelos	42
6.2.1. Politifact-Snopes One Evidence	44
6.2.2. Politifact-Snopes All Evidences	45
6.2.3. News	45
6.3. Evaluación	46
7. Discusión	47
7.1. Explicabilidad de los modelos	47
7.2. Limitaciones de los modelos	47
7.3. Posibles sesgos	48
8. Conclusiones	49
9. Trabajos futuros	51
Bibliografía	51
A. Apéndice	63
A.1. Fragmentos originales	63
A.1.1. Fragmento 1	63
A.1.2. Fragmento 2	63
A.1.3. Fragmento 3	63
A.1.4. Fragmento 4	63
A.1.5. Fragmento 5	63
A.1.6. Fragmento 6	64
A.1.7. Fragmento 7	64
A.2. Visualización SHAP	65
A.2.1. Politifact-Snopes One Evidence	65
A.2.2. Politifact-Snopes All Evidences	67
A.2.3. News	69

Capítulo 1

Introducción

El término *fake news* se origina alrededor del año 1890, aunque diferentes variaciones del concepto se remontan al siglo XVI (Merriam-Webster, 2017). Estas formas de desinformación han ido evolucionando conjuntamente con los medios de transmisión de las sociedades, desde rumores hasta historias completamente fabricadas (Soll, 2016).

Las *fake news* tomaron otra dimensión en las elecciones presidenciales de Estados Unidos en 2016. En los últimos tres meses de campaña electoral las métricas de *engagement* de publicaciones relacionadas con *fake news* superaron a las de los medios tradicionales. Según Dewey (2014), Parkinson (2016) y Read (2016), Donald Trump no habría sido elegido presidente de los Estados Unidos si no fuera por la influencia de las *fake news*.

Uno de los casos más notables en relación a esta industria de las *fake news* se origina en la ciudad de Veles, en Macedonia. Allí se coordinan centenares de páginas web que se dedican a redactar *fake news* que circularán por todo el mundo, aunque gran mayoría están dirigidas al público estadounidense, ya que este es aproximadamente 3 veces más rentable (Subramanian, 2017). Uno de los ‘padres’ de esta industria es Mirko Ceselkoski, que comenzó escribiendo artículos sobre remedios naturales, automóviles y prensa rosa; cambió de un día para otro a escribir *fake news* ya que estas eran aún más lucrativas. Poniéndolo en cifras, se estima que en 2021 aproximadamente un 0.017% de los ingresos mundiales de publicidad acabó en manos de sitios que generan y difunden noticias falsas, lo que equivale a 2.600 millones de dólares (Skibinski, 2022).

Existen dos motivos principales los cuales incentivan la difusión de *fake news*:

- El primer motivo, como se ha dejado entrever anteriormente, es pecuniario: mediante titulares llamativos o historias descabelladas, consiguen llamar la atención del lector, que entra en la página web para conocer más sobre el suceso. Es allí donde a parte de encontrarse con la noticia, se encuentra con una gran cantidad de publicidad, que es la que provee al medio de ingresos, los cuales pueden rondar entre 10.000 y 30.000 dólares mensuales (Sydell, 2016).
- El segundo motivo, ideológico: mediante las *fake news* se busca enturbiar la opinión pública o desacreditar ciertos movimientos, colectivos o personalidades para apoyar su ideología o una agenda política determinada (Allcott & Gentzkow, 2017; Sydell, 2016).

Pero esto no acaba en las páginas web, las redes sociales también sirven de altavoz tanto a las redacciones como a los medios que fabrican noticias falsas. A parte de las motivaciones mencionadas anteriormente, aparecen otros fenómenos propios de estas plataformas.

Las redes sociales poseen de mecanismos los cuales permiten que algunas publicaciones se hagan ‘virales’, permitiendo que estas alcancen un público considerablemente mayor. Gracias al uso de bots que diseminan estas *fake news* o interaccionan con las publicaciones, se genera un aura de legitimidad que provoca que estas sean más creíbles y por consecuencia más difundidas.

Es por ello que el escenario que se dibuja es muy tumultuoso, los medios convencionales y los que difunden *fake news* luchan por encontrar un hueco, mayoritariamente en redes sociales. Según Gottfried y Shearer (2016) y Reid (2017), plataformas como Facebook se han convertido en el medio principal de consumo de noticias en Estados Unidos, aunque esta realidad es extrapolable a gran mayoría de países.

1.1. Motivación

Las *fake news* se han utilizado para manipular la percepción de la realidad de la población, generando desconfianza, y por consecuencia, conflicto social (CITS, University of California Santa Barbara, s.f.). Uno de los casos mas señalados es la teoría conspiratoria *Pizzagate* en Estados Unidos, el Asalto al Capitolio de los Estados Unidos en 2021 y las manifestaciones anti-vacunas en la pandemia de COVID-19 por todo el mundo. Como han demostrado estos sucesos, la desinformación no es para nada inocua en la sociedad: el producto de estos eventos han sido tiroteos, disturbios y fallecidos.

Esta es una de las razones principales por las que las *fake news* dibuja una problemática compleja de solucionar en la sociedad actual: se difunden piezas o fragmentos de información independientemente de su veracidad a una velocidad vertiginosa, la cual dificulta la verificación de esta información. Estas noticias a parte de difundirse rápidamente tienen un alcance mundial, pudiendo afectar a la forma en la que nos relacionamos, comunicamos o directamente al mundo que nos rodea.

Una de las herramientas relevantes para luchar contra la desinformación es el *fact-checking*, realizado mayoritariamente por agencias independientes a los medios de comunicación tradicionales, se encargan de elegir temas de actualidad relevantes y verificar mediante evidencias si un hecho concreto es verídico o no; entre estas agencias se encuentran PolitiFact o Snopes, ubicadas en Estados Unidos, aunque existen homólogos en gran mayoría de países.

El proceso de *fact-checking* requiere de una selección de noticias relevantes, búsqueda de evidencias o recursos, análisis de las fuentes, etc., por lo que es un proceso bastante laborioso y mayoritariamente manual. Es por ello que automatizando ciertas partes puede ayudar a agilizar el proceso de *fact-checking*, pudiendo abarcar mayor variedad o cantidad de noticias a tratar o dedicar más tiempo a tareas que no se pueden desarrollar por métodos automáticos.

Gracias a los *Large Language Models* (LLMs), es posible hacer un pre-análisis de las noticias, desarrollando aplicaciones que permitan ejecutar las siguientes funcionalidades:

- Un clasificador o regresor, que a partir de un fragmento o una noticia completa, sea capaz de predecir si una noticia es verdadera o falsa, o un valor de confianza con respecto a la veracidad de esta, respectivamente.
- Un extractor de palabras clave o conceptos clave, que permita comparar la noticia o un fragmento de esta con otras similares en una base de datos y a partir de este

conjunto de noticias obtenido hacer el análisis manual.

- Una aplicación de resumen automático, que permita obtener un resumen del contenido más importante de la noticia, en forma de lenguaje natural.

Además de la aplicación directa de estas tecnologías para agilizar el proceso de *fact-checking*, creemos que también son relevantes los beneficios que estas aplicaciones pueden proporcionar a la sociedad, como es la reducción de los daños que provocan las fake news, concretamente en colectivos minorizados, que suelen ser el ‘chivo expiatorio’, a parte de perpetuar estereotipos y prejuicios, limitando su pleno desarrollo en sociedad.

1.2. Objetivos

Podemos dividir los objetivos de este trabajo en principales y transversales:

El objetivo principal de este trabajo es de desarrollar una solución que permita clasificar noticias y de esta forma, servir como triaje para las personas encargadas en agencias de *fact-checking*.

Esta clasificación se basará en características estilísticas de los textos, para ello escogeremos dos *datasets* con diferentes características: uno de ellos contendrá todo tipo de recursos estilísticos (palabras capitalizadas, signos de puntuación, etc.), mientras que el segundo estará normalizado. De esta forma podremos determinar en qué medida afectan estos recursos en el rendimiento de los modelos. Otra característica de este último dataset es que contiene diferentes evidencias para la misma noticia o *claim*, por lo que también utilizaremos esto en nuestro análisis, estudiando cómo afecta la cantidad de evidencias en la clasificación.

Para ello, se utilizará una colección de *transformers* basados en la familia de modelos BERT. Estos también serán elegidos de forma que tengamos una gran variedad de tamaños y arquitecturas, permitiéndonos hacer un estudio comparativo de tanto de qué factores, relativos tanto a los modelos como a los datos utilizados, son los que afectan en mayor o menor medida al rendimiento del modelo.

Los objetivos transversales que aparecen al trabajar este tema de investigación son los siguientes:

Un objetivo que aparece al trabajar con los LLMs es conocer en profundidad cómo funcionan estos modelos. Para ello, realizaremos una revisión de la literatura de los trabajos relacionados con el tema de investigación, prestando atención en la casuística del problema, cómo han superado los problemas que han tenido durante el proceso de investigación y diferentes enfoques a la hora de evaluar los resultados obtenidos.

Con respecto al objetivo anteriormente mencionado, es imposible conocer a ciencia cierta como funcionan estos modelos por dentro a la hora de realizar predicciones. Es por ello que aplicaremos técnicas de Explainable AI para intentar entender el ‘razonamiento’ de estos modelos en el proceso de clasificación.

Además, estos modelos al estar entrenados de forma no supervisada con datos generados por humanos, inherentemente presentan sesgos. Conocer estos posibles sesgos y la manera en la los modelos los representan es crucial para el correcto análisis de los resultados, ya que si no se tienen en cuenta, puede desencadenar en resultados incorrectos o incompletos. Para ello, haremos una revisión bibliográfica de diferentes experimentos

realizados para estudiar los efectos de la arquitectura, el proceso de aprendizaje y otras posibles causas en los sesgos aprendidos por estos modelos.

Al estudiar el fenómeno de las *fake news* aparece un problema, el cual es la correcta definición del término. Debido a que este ha sido utilizado constantemente por multitud de medios de comunicación, personas públicas y población general, ha llegado un punto en el que ha perdido su significado. Es por ello que nos proponemos como objetivo definir el término y las posibles implicaciones que pueda tener en nuestro trabajo haciendo una revisión bibliográfica de los trabajos más relevantes relacionados con este cometido. Gracias a este estudio, podremos enfocar nuestro trabajo de forma eficaz, sabiendo qué es lo que realmente queremos investigar y cuáles son las herramientas y recursos de los que disponemos.

1.3. Organización de la memoria

La memoria consta de 8 capítulos:

1. **Introducción.** Presenta el problema de las *fake news* y sus implicaciones en la sociedad actual.
2. **Definiciones.** Define el término *fake news* a partir de trabajos anteriores para definir el problema y el área de acción.
3. **Estado del arte.** Resume los antecedentes, aborda y comenta el estado del arte actual, focalizando el estudio en trabajos con técnicas propias de ciencia de datos.
4. **Análisis del problema.** Analiza el problema a abordar: se presenta el conjunto de datos, describiendo su obtención y procesamiento necesarios; plantea y justifica las técnicas que se utilizarán; y presenta los materiales utilizados para la implementación.
5. **Metodología aplicada.** Describe la metodología aplicada, distinguiendo según las técnicas aplicadas.
6. **Resultados obtenidos y evaluación.** Comenta los resultados obtenidos, haciendo énfasis en las métricas utilizadas y seleccionando las técnicas que mejor rendimiento ofrecen. Discute sobre los resultados obtenidos por técnicas de *Explanable AI*. Finalmente se evalúan ambas partes y se discuten las posibles interpretaciones de los resultados.
7. **Discusión.** Discute ciertos aspectos asociados a la metodología, técnicas y modelos utilizados en el trabajo.
8. **Conclusiones.** Sintetiza los hallazgos más importantes del trabajo.
9. **Trabajos futuros.** Propone ampliaciones al trabajo realizado y propuestas de investigación futuras.

Capítulo 2

Definiciones

Como hemos introducido en el capítulo 1, han habido varios intentos de definir y compartmentar el concepto de *fake news*, ya que de esta manera podemos estudiar de forma íntegra su repercusión. Consideramos interesante incluir estos análisis ya que permite tener una idea de la línea de pensamiento en la literatura, facilitando la comprensión de este fenómeno. Además, ayuda a delimitar el problema, pudiendo trabajar de forma concisa y efectiva, ya que como se detallará a continuación, es un término que estos últimos años ha sido protagonista en los debates de la academia.

Comenzamos con la definición de Allcott y Gentzkow (2017), que es con la cual se ha introducido este trabajo:

artículos de noticias que son intencionada y verificablemente falsos, y que pueden inducir a equívoco a los lectores — [Fragmento original en inglés](#)

La definición de Lazer et al. (2018) también sigue una línea de pensamiento similar, matizando algunos términos:

información fabricada que simula el contenido de los medios de comunicación en la forma, pero no en el proceso organizativo o la intencionalidad — [Fragmento original en inglés](#) —

Estas dos definiciones consiguen dibujar el aspecto de las *fake news*, que consiguen hacerse pasar por artículos periodísticos verídicos y que se caracterizan por tener una intencionalidad de confundir a la población. Además, debido a que no se busca la veracidad, tampoco es necesario seguir los mismos principios organizativos que sigue el periodismo.

El análisis de Tandoc et al. (2017) permite arrojar más luz sobre la intencionalidad, la cual es llegar a que la población legitime tanto estos fragmentos de información como sus autores, alcanzando la misma legitimidad de los medios de comunicación:

las *fake news* se aproximan al aspecto y la esencia de las noticias reales, desde el aspecto de los sitios web hasta la redacción de los artículos o la inclusión de atribuciones en las fotografías. Las *fake news* se esconden bajo un manto de legitimidad, adquiriendo cierta credibilidad al intentar aparecer como noticias reales. Incluso, yendo más allá de la simple apariencia de una noticia, mediante el uso de *bots*, las *fake news* imitan la omnipresencia de las noticias construyendo una red de sitios falsos. — [Fragmento original en inglés](#)

Aún así, estas definiciones son insuficientes, ya que según un estudio de Mourão y Robertson (2019), se encontraron entre la gran mayoría de los fragmentos analizados una mezcla de información falsa, sensacionalismo, contenido sesgado y *clickbait*¹. A esto se suma el hecho de que no todo el contenido falso que se difunde tiene estructura o apariencia de noticia. Por tanto, es crucial reformular el concepto, ya que no nos permitiría afrontar o incluso entender el problema en su totalidad.

Si consideramos todas las formas de información falsa en Internet como noticias falsas cuando estas no se presentan en formato de noticia, estamos atentando contra el rigor que se espera de una investigación calidad (MacKenzie et al., 2011; Suddaby, 2010; M. Zhang et al., 2016). Por otro lado, excluir estas formas de falsedad por no tener carácter de noticia podría mermar su relevancia al pasar por alto cuestiones del mundo real, como la negación del cambio climático y el llamado movimiento anti-vacunas, que se nutren de investigaciones, informes o publicidad fraudulentos o cuestionables, y de opiniones partidistas (Khan et al., 2021). Esto incluso puede desencadenar en errores de ‘tipo III’ en la formulación de problemas de investigación: el hecho de centrarse en la cuestión inmediata sin tener en cuenta “un problema más general y arquetípico”, impediría hacer una “contribución académica más amplia y duradera a escala del problema genérico” (Rai, 2017).

Es por ello que Khan et al. (2021) formula el problema utilizando como referencia el concepto de información de Mingers y Standing (2018): “el contenido proposicional de signos”.

La información es un contenido proposicional puesto que propone la existencia de un estado concreto del mundo, ‘lo que debe ocurrir en el mundo para que el signo exista como y cuando existe’. — [Fragmento original en inglés](#)

Y a partir del concepto de información formula los términos *misinformation*, *disinformation* y *malinformation*:

Misinformation. Contenido proposicional de signos que tergiversa el estado del mundo sin intención de engañar [...]. Un área [...] en la (este fenómeno) es bastante común es el asesoramiento sanitario en comunidades en línea (Venkatesan et al., 2014), donde muchas personas difunden información falsa de forma no intencionada (Myers & Pineda, 2009). (Khan et al., 2021) — [Fragmento original en inglés](#)

Disinformation. Contenido proposicional de signos que tergiversa el estado del mundo con la intención de engañar. (Khan et al., 2021) — [Fragmento original en inglés](#)

Malinformation. Contenido proposicional de signos que representa verazmente el estado del mundo con la intención de engañar [...] A menudo se asume que el engaño aparece en forma o como resultado de la mentira descarada y otras formas de *disinformation*. Sin embargo, el engaño puede producirse igualmente en forma o como resultado de una sutil manipulación de la información

¹Entendemos *clickbait* como el diseño de contenidos con el objetivo de llamar la atención de los lectores y atraerlos para que hagan clic en un enlace a un sitio web mediante tácticas como historias sensacionalistas, titulares llamativos e imágenes, que funcionan como cebo (Blom & Hansen, 2015; Y. Chen et al., 2015)

que no necesariamente tergiversa el mundo pero que tiene la intención de engañar (McCornack, 2009; McCornack et al., 2014; Wardle, 2018b). Algunos ejemplos son las medias verdades y los montajes, que se refieren a información incompleta o selectiva proporcionada con la intención de engañar (Fallis, 2016). (Khan et al., 2021) — Fragmento original en inglés

Habiendo estudiado exhaustivamente qué constituyen las *fake news* y hasta dónde abarcan, podemos establecer concretamente el ámbito de aplicación de nuestro estudio. Para ello, tomaremos la definición de disinformation: contenido proposicional de signos que tergiversa el estado del mundo con la intención de engañar.

Debido a que la tipología del problema es compleja, limitaremos nuestra área de aplicación a noticias, ya que estas son las más fáciles de recopilar gracias a repositorios de páginas web catalogadas como *fake news*. Asimismo, estas son fácilmente categorizables como verdaderas o falsas, facilitando el entrenamiento de los modelos, que se hará de manera supervisada.

Capítulo 3

Estado del arte

3.1. Antecedentes

El problema de la detección automática de *fake news* no se ha conceptualizado hasta la publicación de los trabajos de Cohen et al. (2011) y Flew et al. (2012). Gracias a los avances en procesado del lenguaje natural, bases de datos e *information retrieval* de la época, la idea es poder ayudar a los periodistas con su tarea de *fact checking* o incluso actualizar en tiempo real estadísticas o hechos automáticamente a medida que van sucediendo, de forma que el lector siempre tenga la última información relevante cuando acceda al artículo.

Los cambios en las tendencias de consumo producidos por la masificación de Internet también han afectado al consumo de noticias, tanto en la cantidad como en la personalización que medios como las redes sociales o algoritmos de *profiling* pueden ofrecer. Gracias a que las redes sociales permiten tener la misma facilidad de acceso sumado a la personalización del contenido ofrecido, los usuarios dependen más de estas para consumir noticias.

Como hemos comentado en el capítulo 1, redes sociales como Facebook se han convertido en el medio principal de consumo de noticias en Estados Unidos (Gottfried & Shearer, 2016; Reid, 2017). En Australia, alrededor del 60 % de los usuarios de noticias online se pueden categorizar como ‘de conveniencia’, por lo que su fuente de contenido o noticias provendrá de la red social y el medio que más se ajuste a las necesidades del lector (Daniel et al., 2009).

El periodismo de investigación es un proceso muy tedioso, costoso y lento. Este proceso implica consultar fuentes en infinidad de formatos (archivos excel, PDF, páginas web, vídeo, foto, audio, etc.), por lo que es muy complicado de estandarizar. Aunque este proceso no parece ni resulta rentable, es necesario para los medios como forma de mantener una imagen de marca que genere credibilidad.

Hay dos enfoques a este problema: basado en patrones y basado en evidencias.

Enfoque basado en patrones. Los modelos utilizan solamente consideran el estilo o la sintaxis del texto. Popat et al. (2016) implementó el modelo basándose en *features* estilísticas y la postura general del artículo. Otros trabajos se aprovechan de las métricas que proporcionan las redes sociales (*likes*, veces compartido, comentarios, etc. para determinar la ‘veracidad’ de una cuenta o medio (Ajao et al., 2019; Benamira et al., 2019; Q. Liu et al., 2017; Popat et al., 2016; Vo & Lee, 2018; Volkova et al., 2017; Yu et al., 2017). Por último con respecto a minería de datos enfocada a emociones, los trabajos de Ajao

et al. (2019), Giachanou et al. (2019) y X. Zhang et al. (2019) muestran que es viable su implementación.

Enfoque basado en evidencias. Se basan en explorar la similaridad semántica entre *claim* y *evidencia*. Estas evidencias se obtienen a partir de un grafo de conocimiento o páginas de *fact checking* utilizando las *claims* como término de búsqueda. El trabajo de Popat et al. (2018) es el primero en utilizar evidencias para clasificación de noticias, Ma et al. (2019) y Wu, Rao, Yang et al. (2021) lo precede.

3.2. Tecnologías

Los trabajos más notables con respecto a esta tarea son los siguientes:

DeClarE (Popat et al., 2018). Utiliza evidencias y contraevidencias obtenidas de Internet para apoyar o refutar una afirmación. El modelo consta de tres componentes principales: un verificador, un extractor de evidencias y un clasificador de afirmaciones.

El modelo utiliza una combinación de redes neuronales convolucionales (CNN) y redes neuronales recurrentes (RNN) para extraer características de las evidencias y contraevidencias. Las CNN se utilizan para extraer características locales de cada pieza de evidencia, mientras que las RNN se utilizan para capturar las dependencias globales entre diferentes piezas de evidencia. El modelo también utiliza mecanismos de atención para aprender qué piezas de evidencia son más importantes para determinar la veracidad de la afirmación.

HAN (Ma et al., 2019). Explota el proceso de extracción de *features* mediante una *Hierarchical Attention Network* (HAN) y las características visuales de la imagen utilizando *image captioning* y análisis forense. El modelo consta de tres partes principales: codificador, decodificador y detector de noticias falsas. El autoencoder variacional está equipado para aprender modelos de variables inactivas probabilísticas mediante optimización. HAN tiene dos niveles de mecanismos de atención uno a nivel de palabra y otro de oración, lo que le permite atender diferencialmente al contenido más y menos importante al construir la representación del texto verdadero/falso.

GET (Xu et al., 2022). Propone un marco unificado de minería de la estructura semántica de texto mediante grafos, gracias a ellos se puede capturar la dependencia semántica a larga distancia entre fragmentos relevantes, los cuales no suelen ser percibidos por los modelos por esta razón. Después de esta información, el modelo refina el grafo. Finalmente, las representaciones semánticas detalladas se alimentan al módulo de interacción afirmación-evidencia para hacer predicciones.

El modelo GET se compone de cuatro componentes principales: *Graph Construction*, *Graph-based Semantics Encoder*, *Semantic Structure Refinement* y *Attentive Graph Readout Layer*.

El módulo de *Graph Construction* construye un grafo para cada afirmación y sus correspondientes evidencias o pruebas. Los nodos en el grafo representan las palabras en las afirmaciones y pruebas, y las aristas representan sus relaciones semánticas. Después, utilizan un modelo de lenguaje pre-entrenado para obtener *word embeddings* contextualizados de cada nodo.

El módulo *Graph-based Semantics Encoder* codifica las *word embeddings* en un espacio de baja dimensionalidad utilizando una red convolucional basada en grafos (GCN).

El módulo *Semantic Structure Refinement* reduce la redundancia de información realizando el aprendizaje de la estructura del grafo. Captura la dependencia semántica a larga distancia entre fragmentos relevantes dispersos a través de la propagación del vecindario.

Finalmente, el módulo *Attentive Graph Readout Layer* captura las interacciones entre afirmaciones y pruebas utilizando un mecanismo de atención *multi-head*.

Capítulo 4

Análisis del problema

4.1. Estudio preliminar

En un principio queríamos analizar el discurso de odio y las *fake news* que se difunden por redes sociales, motivado por los trabajos de Gomez et al. (2019) y Toraman et al. (2022).

Algunos de los *datasets* considerados fueron los siguientes:

- VoterFraud2020 (Abilov et al., 2021): *Dataset* multimodal de *tweets* y *retweets* en inglés que contienen palabras clave y *hashtags* relacionados con alegaciones y reclamaciones sobre fraude electoral en las elecciones presidenciales de Estados Unidos en 2020.
- Hate Speech Corpus: *Dataset* en inglés obtenido a partir del trabajo realizado por Szpakowski (2017). Extraen artículos de medios catalogados por publicar contenido de discurso de odio en Internet. Estos artículos tratan anti-semitismo, misoginia, anti-inmigración/xenofobia, homofobia o discurso de odio en general.
- MMHS150K (Gomez et al., 2019): *Dataset* multimodal en inglés que incluye *tweets* e imágenes relacionadas con discurso de odio.
- Hate + COVID + Temporers (Rodríguez et al., 2022): *Dataset* de *tweets* en castellano y catalán relacionados con COVID-19 y temporeros en Cataluña, los cuales la mayoría de ellos suelen ser migrantes o pertenecen a poblaciones minorizadas. Los *tweets* se recogieron en el periodo de enero a octubre del 2020, obteniendo un total de 1.062 *tweets*.

Debido a que las pruebas de LLMs con estos *datasets* no obtenían resultados satisfactorios, decidimos redirigir la investigación hacia detección de noticias falsas, las cuales tienen estrecha correlación con el discurso de odio.

Una de las razones por las que creemos que los LLMs no funcionaban con estos *datasets* era por la poca cantidad de texto que contenían en los casos de Abilov et al. (2021), Gomez et al. (2019) y Rodríguez et al. (2022). Esto se justifica con el hecho de que los datos obtenidos son procedentes de Twitter y en el momento de recolección de los datos existía un límite de 280 caracteres por *tweet*. Al no tener suficiente información de texto por *tweet* y por consecuencia no tener suficiente contexto, suponemos que los LLMs no eran capaces de detectar las características necesarias para realizar la clasificación.

Además, los *datasets* de Abilov et al. (2021) y Szpakowski (2017) no tenían casos negativos (es decir, muestras que no estuvieran categorizadas como discurso de odio). Esto suponía realizar una creación de un *dataset* de casos negativos, por lo que decidimos finalmente descartar estos *datasets*.

4.2. Datasets

4.2.1. Descripción del conjunto de datos

Politifact y Snopes

Snopes (Popat et al., 2017) y PolitiFact (Vlachos & Riedel, 2014) son dos *datasets* creados a partir de las agencias de *fact-checking* homónimas.

Ambos *datasets* son obtenidos gracias a Xu et al. (2022), que utilizan ambas páginas para obtener las *claims* y etiquetas para cada noticia (en el caso de Snopes son *True/False*). A partir de cada *claim*, obtienen las evidencias y demás información mediante motores de búsqueda. Para el caso de PolitiFact la única diferencia que existe es que se fusionan las clases positivas (*true*, *mostly true* y *half true*) en la categoría *true*, mientras que las negativas (*false*, *mostly false* y *pants on fire*) como *false*. Un último dato relevante sobre este dataset es que está normalizado, por lo que los modelos entrenados con este dataset tendrán solo acceso a la información contextual o estilística de las noticias.

En la figura 4.1 se pueden encontrar diversas estadísticas sobre el dataset:

Dataset	Feature	Conteo
PolitiFact	True	1543
	False	1565
	Evidences	3108
	Speakers	137
	Publishers	1064
Snopes	True	690
	False	2066
	Evidences	2756
	Speakers	N/A
	Publishers	1873

Cuadro 4.1: Conteo de muestras según *features*.

ISOT Fake News Dataset

Este conjunto de datos contiene artículos periodísticos reales y falsos. Los artículos reales fueron obtenidos de la agencia de noticias Reuters, mientras que los artículos falsos han sido obtenidos de medios poco fiables catalogados por PolitiFact (una agencia de verificación de noticias de EE.UU.) y Wikipedia.

Los temas que incluye este *dataset* abarcan diferentes ámbitos, aunque como se puede observar en la figura 4.2, la gran mayoría tratan sobre política y actualidad mundial.

News	Subjects		Total size
	Type	Size	
Real-News	World-News	10145	21417
	Politics-News	11272	
Fake-News	Government-News	1570	23481
	Middle-East	778	
	US News	783	
	Left-News	4459	
	Politics	6841	
	News	9050	

Cuadro 4.2: Conteo de muestras según categorías (Ahmed et al., 2017).

Cada artículo contiene la siguiente información: título del artículo, texto, tipología y fecha de publicación. Las noticias recogidas en este *dataset* están limitadas al periodo entre 2016 y 2017. Posteriormente, estas noticias fueron limpiadas y procesadas, aunque en el caso de las noticias falsas no se corrigieron errores tipográficos, en los cuales se incluyen errores ortográficos o de puntuación.

Este *dataset* mantiene un equilibrio entre clases, por lo que no es necesario aplicar técnicas de *Data Augmentation*.

4.2.2. Procesamiento del conjunto de datos

Politifact y Snopes

Para trabajar con ambos *datasets* fusionaremos ambos *datasets*, creando una nueva columna **dataset** que indica si pertenece al *dataset* **politifact/snopes** (por si es necesario en un futuro análisis). Concatenamos los *strings* para el titular y cuerpo de la noticia.

Debido a que contamos con varias evidencias para el mismo *claim*, generaremos dos *datasets* a partir de este, P-S_{One} y P-S_{All}. La diferencia entre ambos es que el primero de ellos contendrá solamente una evidencia, elegida aleatoriamente; el otro contendrá todas las evidencias concatenadas junto al titular, como hemos mencionado anteriormente.

De esta forma, ambos *datasets* P-S_{One} y P-S_{All} tendrán el mismo número de muestras, 789, aunque con diferente información, teniendo el *dataset* P-S_{All} más información contextual sobre el hecho sucedido.

De esta forma, obtenemos estos dos *datasets* con las siguientes características:

Dataset	Etiqueta	Conteo	Total
PolitiFact	True	186	356
	False	170	
Snopes	True	116	433
	False	317	

Cuadro 4.3: Conteo de muestras según categorías.

Podemos observar que existe un pequeño desbalanceo entre clases *True/False*, donde la proporción entre noticias falsas y verdaderas es de 3:2 respectivamente. Debido a que

este desbalanceo no es tan pronunciado, consideramos que no es necesario aplicar alguna técnica de balanceo entre clases.

Por último, con respecto al *dataset* P-S_{All}, consideramos pertinente mostrar las siguientes estadísticas sobre el número de evidencias por noticia:

Estadístico	Valor
Media	7,43
Desviación estándar	6,34
Mínimo	1
Q ₁	2
Q ₂	5
Q ₃	11
Máximo	27

Cuadro 4.4: Estadísticas con respecto al número de evidencias por noticia.

Como podemos observar, al menos el 75 % de noticias tienen 2 evidencias o más, por lo que aunque los *datasets* P-S_{One} y P-S_{All} parten del mismo conjunto de datos original, P-S_{All} tiene considerablemente más información contenida por noticia que su contraparte P-S_{One}.

ISOT Fake News Dataset

En este *dataset* podemos encontrar para las noticias reales que todas comienzan por la ubicación del suceso y la agencia de noticias, de la siguiente forma:

‘BARCELONA/GIRONA, „Spain“ (Reuters) –’

También es posible encontrar aclaraciones a principio de esta información, como puede ser:

‘(This „Oct. 9 story has been refiled to add a dropped word in the headline) By Sonya Dowsett’

Es por ello que hemos borrado estos fragmentos de texto para evitar que el modelo pueda aprender a utilizar estas estructuras y diferenciar entre noticias verdaderas y falsas, ya que en el caso de las falsas esto no sucede.

Como el *dataset* está dividido en dos archivos, cada uno contenido las noticias de una label concreta (*True/False*), añadimos una columna adicional para codificar la categoría a la que pertenecen y concatenamos ambos *datasets* para obtener el *dataset* final con todas las noticias.

Posteriormente, concatenamos los *strings* para el titular y el cuerpo de la noticia.

Normalización y limpieza

Después de haber obtenido los *datasets* procesados con los que vamos a entrenar los modelos, aplicamos un paso adicional de normalización y limpieza de texto para los modelos *Bag of Words* y TF-IDF.

Para ello, aplicamos los siguientes pasos:

1. Sustituimos diferentes acrónimos de estados, organizaciones, siglas y palabras malsonantes censuradas por sus equivalentes.
2. Convertimos el texto a minúsculas
3. Expandimos contracciones
4. Eliminamos enlaces web, *tags* HTML, caracteres mal codificados, símbolos de puntuación y otros caracteres indeseados.
5. Eliminamos *stop words*
6. Tokenizamos cada palabra
7. Lematizamos cada token

Creación de conjuntos de entrenamiento, validación y test

Para el entrenamiento de los modelos *Bag of Words* y TF-IDF, hacemos uso de la librería `scikit-learn` (Pedregosa et al., 2011) y hacemos dos conjuntos: entrenamiento y test, debido a que no es necesario un conjunto de validación en esta implementación.

Estos dos conjuntos se dividen en un ratio 8:2 para entrenamiento y test respectivamente. De esta manera, la misma proporción de muestras se dedican al entrenamiento para estos modelos como para los *transformers*.

En el caso del entrenamiento de los modelos basados en *transformers*, aprovechamos la clase `Dataset` de la librería `datasets` (Lhoest et al., 2021) y creamos tres conjuntos: entrenamiento, validación y test. Como hemos adelantado, la división se hace de forma 8:1:1 respectivamente; evitando que ninguno de los modelos tengan acceso a un mayor número de muestras durante su entrenamiento.

4.3. Modelos y clasificadores utilizados

4.3.1. Bag of Words

Es una técnica de representación de documentos la cual se basa en contabilizar el número de veces que aparece cierta palabra en el documento. De esta forma, dos documentos serán similares si contienen las mismas palabras (Vajjala et al., 2020). Este modelo es denominado de esta manera ya que solamente incluye información sobre el conteo de cada palabra, ignorando cualquier información (gramatical, contextual, etc.) a excepción de las palabras en si (Eisenstein, 2019).

4.3.2. TF-IDF

TF-IDF modela la información de forma similar al modelo *Bag of Words*, ya que no almacena ningún tipo de información gramatical o contextual. Lo novedoso de este modelo es que introduce los conceptos de *term-frequency* y *inverse document frequency*.

Se basa en el siguiente hecho: si una palabra concreta aparece repetidas veces en un documento pero no en el resto, entonces esta palabra debe ser importante o representativa

para ese documento. Es por ello que la importancia de una palabra determinada debe incrementarse proporcionalmente a su frecuencia en un documento determinado; también, esta debe disminuir proporcionalmente a su frecuencia en otros documentos del corpus (Vajjala et al., 2020).

Matemáticamente, esta relación es capturada usando la *term-frequency* y *inverse document frequency*, que combinadas generan el *TF-IDF score*.

TF (term-frequency) mide la frecuencia de una palabra p en un documento d .

IDF (inverse document frequency) mide la importancia de una palabra p en el corpus, dando más peso a las palabras menos frecuentes (o las más inusuales). Se define como:

$$\text{IDF}(p, D) = \ln \frac{D}{|\{d \in D : p \in d\}|} \quad (4.1)$$

Siendo D el número total de documentos en el corpus y $|\{d \in D : p \in d\}|$ el número de documentos donde la palabra p aparece.

Finalmente el *TF-IDF score* se calcula de la siguiente manera:

$$\text{TF-IDF}(p, d, D) = \text{TF}(p, d) \cdot \text{IDF}(p, D) \quad (4.2)$$

4.3.3. Regresión logística

La regresión logística es un método estadístico que modela la relación entre una variable dependiente y una o más variables independientes. Se utiliza para problemas de clasificación binaria.

Para tareas de clasificación de texto, la regresión logística funciona calculando una suma de las características de entrada (en nuestro caso la representación mediante *Bag of Words* o TF-IDF) y calculando mediante una función logística el resultado.

El modelo de regresión logística toma los valores de entrada y aprende a predecir la probabilidad de cada clase basándose en las características de entrada. El modelo hace esto aprendiendo un conjunto de pesos para cada característica que se utilizan para calcular una suma ponderada de las características de entrada. Esta suma ponderada se pasa luego a través de la función logística para producir un valor de probabilidad para cada clase.

4.3.4. Naïve Bayes

Naïve Bayes es un algoritmo probabilístico basado en el teorema de Bayes. Establece que la probabilidad de una hipótesis (en este caso, una clase) es proporcional a la probabilidad de la evidencia (en este caso, las características de entrada) dada esa hipótesis.

En la clasificación de texto, Naïve Bayes funciona calculando la probabilidad de cada clase dada las características de entrada. Lo hace asumiendo que las características de entrada son condicionalmente independientes dadas las clases. Esta suposición se llama suposición ‘ingenua’ o ‘naïve’, y es lo que le da al algoritmo su nombre.

El algoritmo Naïve Bayes luego calcula la probabilidad de cada clase dada las características de entrada utilizando el teorema de Bayes. Específicamente, calcula la probabilidad previa de cada clase (es decir, la probabilidad de cada clase antes de ver ninguna

evidencia), y luego multiplica esto por la verosimilitud de la evidencia dada cada clase (es decir, la probabilidad de ver las características de entrada dadas cada clase). Finalmente, normaliza estas probabilidades para obtener una distribución de probabilidad sobre las clases.

4.3.5. SVM y SGD

Estos métodos supervisados funcionan aprendiendo una función que mapea las características de entrada a las clases de salida. Por esta razón, son similares a la regresión logística.

Las máquinas de vectores de soporte (SVM) son un tipo de clasificador que funcionan encontrando el hiperplano que separa al máximo las características de entrada en diferentes clases. El hiperplano se elige de manera que maximice el margen entre las clases. El margen se define como la distancia entre el hiperplano y los puntos de datos más cercanos de cada clase.

La función de pérdida del perceptrón es un tipo de función de pérdida que se utiliza para entrenar SVM. Funciona minimizando la distancia entre la salida predicha y la salida verdadera. Específicamente, minimiza la suma de las distancias entre cada salida predicha y su correspondiente salida verdadera.

4.3.6. Random Forest

El algoritmo Random Forest funciona construyendo múltiples árboles de decisión y luego combinando sus predicciones.

Cada árbol de decisión se construye seleccionando aleatoriamente un subconjunto de las características de entrada y luego particionando recursivamente los datos en función de estas características. La partición se realiza de tal manera que los subconjuntos resultantes sean lo más puros posible con respecto a las clases de salida.

Una vez que se han construido todos los árboles de decisión, sus predicciones se combinan para obtener una predicción final. Esto se hace típicamente tomando una mayoría de votos sobre todos los árboles de decisión.

4.3.7. BERT

Devlin et al. (2018) discute que las técnicas de aprendizaje hasta el momento limitan las capacidades de los LLMs, concretamente ELMo (Peters et al., 2018) y GPT (Radford et al., 2018) tienen una forma de aprendizaje unidireccional. Esto puede mermar tareas como *question answering*, las cuales se benefician de disponer de contexto por ambos lados.

BERT está basado en la arquitectura de los Transformers propuesta por Vaswani et al. (2017) y propone una arquitectura bi-direccional en la cual el modelo aprenderá siguiendo estas dos tareas:

Inspirados en la Prueba cloze o *Cloze test* (Taylor, 1953), se enmascaran una parte de los *tokens* en el texto para que el modelo prediga cuál va en su lugar; esta tarea la denominan como *Masked LM* (MLM). Esta tarea se distingue de los *denoising auto-*

encoders en el hecho de que el objetivo consiste en predecir el *token* enmascarado, en lugar de reconstruir completamente el *input*.

La segunda tarea que desempeña el modelo en su entrenamiento es la denominada *Next Sentence Prediction* (NSP). A partir de un par de frases, el modelo debe clasificar si la segunda frase es continuación de la primera.

4.3.8. DistilBERT

DistilBERT (Sanh et al., 2019) se basa en el concepto de *Knowledge Distillation*, donde un modelo compacto ‘alumno’ es entrenado para reproducir el comportamiento del modelo ‘profesor’, de mayor tamaño.

Siguiendo una arquitectura del modelo ‘profesor’, el modelo ‘alumno’ tiene una arquitectura similar aunque reducida en el número de capas, el cual se reduce a la mitad.

Otro truco se aplica en la inicialización de los parámetros del modelo ‘alumno’: aprovechando que la arquitectura de ambos modelos es similar y por lo tanto su dimensionalidad también es similar, aprovechan los pesos del modelo ‘profesor’ para inicializar los pesos del ‘alumno’.

En resumen, DistilBERT consigue reducir el tamaño del modelo en un 40% y la complejidad de computación en un 60%, reteniendo el 97% de capacidades del modelo original.

4.3.9. RoBERTa

RoBERTa (Y. Liu et al., 2019) mejora el rendimiento de BERT mediante mejoras en el proceso de aprendizaje y mejora de hiperparámetros, ya que descubrieron que este estaba infra-entrenado. Es por ello que proponen los siguientes cambios con respecto a la arquitectura y el proceso de entrenamiento:

Primero, descartan el entrenamiento mediante *Next Sentence Prediction*, ya que consideran que su contribución mina el rendimiento global del modelo. Es por ello, que después de diversos experimentos llegan a la conclusión que eliminando esta tarea se consiguen rendimientos iguales o mejores, por lo que se puede prescindir perfectamente de esta tarea.

La siguiente modificación tiene relación con el valor del *learning rate* y el *batch size*. El trabajo de Ott et al. (2018) dejan entrever que aumentar el *batch size* permite una aceleración en el aprendizaje además de una mejora en el rendimiento del modelo, siempre y cuando el valor del *learning rate* se ajuste acordemente. Por otro lado, también se ha descubierto que BERT se puede aprovechar de este tipo de entrenamiento (You et al., 2019).

Motivado por los trabajos mencionados anteriormente, Y. Liu et al. (2019) aumenta *batch size* de 256 llegando hasta 8.192 muestras por *batch*, obteniendo mejoras en los valores de *perplexity* y rendimiento en tareas finales.

Por último, la última mejora propuesta se enlaza con la tarea de *Masked LM*: se mejora el proceso de enmascaramiento de los *tokens*: en el caso de BERT este enmascaramiento es estático, enmascarando solo un tipo determinado de *tokens*. El enmascaramiento dinámico de RoBERTa enmascara diferentes tipos de *tokens* según cada época en el entrenamiento, haciendo que el modelo sea menos dependiente en los patrones de las frases y por tanto

más robusto.

Después de comentar los diferentes aspectos con respecto al proceso de entrenamiento, solamente quedaría mencionar el corpus utilizado. Inspirado por el trabajo de Yang et al. (2019), decidieron utilizar también un corpus considerablemente grande, sobre unas 10 veces más grande que el utilizado para entrenar BERT. Esto resulta ser idóneo para obtener un buen rendimiento en el modelo sin dar indicios de sobreajuste.

4.3.10. DeBERTa

DeBERTa (He et al., 2020) propone una mejora en la arquitectura de BERT (Devlin et al., 2018) la cual se basa en el principio de *Disentangled Attention*:

En el caso de BERT, cada palabra en la capa de *inputs* es representada en forma de vector como la suma del valor del *word embedding* y *position embedding*. DeBERTa propone trabajar estos dos valores por separado y calcular los valores de atención mediante *disentangled matrices*, teniendo en cuenta estos valores de posición y contenido.

Esto se justifica en el hallazgo de que la atención para un par de palabras depende tanto del contenido de las palabras como de su posición relativa entre ellas.

Mediante este novedoso mecanismo de atención, DeBERTa consigue mejorar el rendimiento de BERT y RoBERTa, especialmente en tareas que requieren un razonamiento exhaustivo de diferentes partes de la *input*.

4.4. Material, recursos utilizados e implementación

El trabajo realizado en este proyecto se ha desarrollado en Python 3.9 (Van Rossum & Drake, 2009), utilizando Jupyter Notebook (Kluyver et al., 2016) como interfaz de programación principal para la implementación. El código ha sido desarrollado y probado a pequeña escala en un ordenador de tipo usuario, mientras que el grueso del entrenamiento, evaluación y testeo de los modelos ha sido ejecutado en un servidor proporcionado por la Universitat de València.

Las librerías utilizadas fueron las siguientes: **pandas** (McKinney, 2010), **numpy** (Harris et al., 2020), **scikit-learn** (Pedregosa et al., 2011), **nltk** (Bird et al., 2009), **pytorch** (Paszke et al., 2019), **transformers** (Wolf et al., 2019), **datasets** (Lhoest et al., 2021), **shap** (S. M. Lundberg & Lee, 2017).

Por otro lado, las especificaciones del servidor provisto son las siguientes:

- 1x Intel® Xeon® CPU E5-2650 v4 @ 2.20GHz
- 264 GB RAM DRR4
- 1x NVIDIA® Tesla® P100-PCIE-12GB

El *script* de entrenamiento de los modelos tardó en ejecutar un total de 2 días y 13 horas de forma ininterrumpida, ya que se ejecutó con el comando **nohup** (Brady, 2017). El resto del código implementado ha sido ejecutado de forma *online* con Jupyter Notebook debido a las facilidades que aporta.

Capítulo 5

Metodología aplicada

5.1. Entrenamiento de modelos estadísticos

Trabajaremos con la librería `scikit-learn` (Pedregosa et al., 2011), la cual tiene gran variedad de modelos y clasificadores. Mediante un bucle, seleccionamos en cada iteración un vectorizador (BoW/TF-IDF), que se entrenará con el conjunto de entrenamiento.

Los parámetros elegidos para los vectorizadores son los siguientes:

- `analyzer='word'`: se encarga de hacer la división entre *tokens*, en este caso, la unidad fundamental será la palabra.
- `ngram_range=(1,2)`: contabilizamos los tokens tanto individualmente como en combinaciones de dos (bigrama).
- `token_pattern=r'\w{1,}'`: descarta los tokens que no sean caracteres, dígitos ni ‘_’.
- `min_df=2`: umbral inferior para eliminar ocurrencias, en este caso, unigramas o bigramas que no aparezcan en al menos dos documentos no serán registrados.

Para los clasificadores, estos son los parámetros elegidos:

- `max_iter=10000`: número máximo de épocas que se ejecuta el entrenamiento.
- `tol=1e-5`: criterio de parada, similar a *early stopping*; si el valor para la *training loss* es menor que este valor se detiene el entrenamiento.
- `random_state=42`: *seed* para fijar el entrenamiento y no obtener resultados diferentes entre ejecuciones. Solamente sirve en algoritmos heurísticos.

Después, realizamos la predicción con el conjunto de test, guardamos los resultados y calculamos las métricas.

5.2. Entrenamiento de transformers

Hacemos uso de la librería `transformers` (Wolf et al., 2019) para agilizar el entrenamiento de todos los modelos gracias a las clases `AutoTokenizer` y `AutoModelForSequenceClassification`, que permite importar los *tokenizers* y los modelos de lenguaje fácilmente

para una serie de modelos que han sido configurado por los desarrolladores previamente para este fin.

Para cada modelo utilizamos su *tokenizer* correspondiente, añadiendo padding o truncando la *input* si es necesario.

Seleccionamos un *batch size* lo suficientemente grande para acelerar el tiempo de entrenamiento sin desbordar la memoria de la tarjeta gráfica. Existen librerías para determinar programáticamente estos valores, pero por practicidad decidimos determinarlo a mano, obteniendo unos valores de *batch size* entre 32 y 2 muestras dependiendo del tamaño del modelo.

El valor del *learning rate* lo establecemos a $2e-5$ con un valor de *weight decay* de $1e-2$.

Aunque en este trabajo no se haya llegado a implementar un mecanismo de *early stopping*, establecemos la estrategia de evaluación cada época para supervisar que los modelos aprenden correctamente.

Establecemos también un valor de *seed* para fijar los resultados obtenidos en diferentes iteraciones,

Finalmente, para poder generalizar los resultados entre modelos, decidimos establecer el número de épocas de entrenamiento a 3. Esto nos proporciona un equilibrio entre tiempo de ejecución necesario para los modelos y un aprendizaje suficiente para poder obtener unos resultados coherentes. También fijamos un valor de *seed* y así evitar obtener resultados diferentes entre ejecuciones.

Después del entrenamiento con la librería `transformers` (Wolf et al., 2019), hacemos la predicción utilizando el conjunto de test. En este caso, por facilidad y flexibilidad a la hora de programar, utilizamos `pytorch` (Paszke et al., 2019). Para cada modelo y *dataset* generamos las predicciones y guardamos los resultados para su futuro análisis.

Capítulo 6

Resultados obtenidos y evaluación

6.1. Resultados obtenidos

Los resultados de precisión para todos los modelos y *datasets* utilizados se muestran en la tabla 6.1. Observamos que los modelos *Bag of Words* (BoW) y TF-IDF obtienen el mejor rendimiento (con respecto a *accuracy*) para los tres *datasets*. Esto es más notable en el caso del *dataset* News, el cual el hecho de pasar a un modelo basado en *transformers* supone una pérdida de 0,5 puntos aproximadamente con respecto a precisión.

En los casos donde se utilizan los *datasets* PolitiFact-Snopes con una evidencia (P-S_{One}) y con todas las evidencias (P-S_{All}), observamos un decrecimiento en el rendimiento para los modelos DistilBERT (exceptuando DistilBERT_{BASE}, CASED, MULTILINGUAL), BERT_{BASE} y RoBERTa_{BASE}. Para los modelos BoW y TF-IDF podemos observar una mínima ganancia de rendimiento de 0,2 y 0,7. En los modelos BERT_{LARGE}, RoBERTa_{LARGE} y DeBERTa, el rendimiento se mantiene en su línea.

Los resultados de *precision*, *specificity* y *F1* de la clase negativa, *fake news* en todos los modelos y *datasets* utilizados podemos encontrarlos en el cuadro 6.2. Elegimos calcular las métricas de esta manera debido a que creemos que es la más representativa para detectar *fake news*:

- La *precision* o *Positive Predictive Value* nos indica cómo de eficaz es a la hora de detectar los verdaderos positivos, es decir, de las veces que el modelo predice que una noticia es *fake news*, cuántas realmente lo son.
- La *specificity* o *True Negative Rate* nos indica como de eficaz es el modelo de detectar los verdaderos negativos, de todas las noticias verdaderas, cuántas de ellas han sido predichas como tal.
- El *F1-Score* nos indica como de equilibrado es el clasificador con respecto a *precision* y *specificity*.

En líneas generales, observamos que los modelos BoW y TF-IDF junto a los LARGE basados en *transformers* son los que ofrecen unas métricas razonables con respecto a *precision*, *specificity* y *F1*, aunque procederemos a comentar métrica por métrica qué modelos funcionan mejor y las implicaciones de cada métrica en su valoración.

Teniendo en cuenta el valor de *precision*, observamos que obtienen las mejores métricas en los *datasets* de P-S_{One}, P-S_{All} y News los modelos BoW + SGD, BoW + NB y

Model	P-S _{One}	P-S _{All}	News
BoW + LR	0,65	0,66	0,99
BoW + NB	0,63	<u>0,70</u>	0,97
BoW + SVM	0,63	0,68	0,98
BoW + SGD	<u>0,66</u>	0,68	0,99
BoW + RF	0,63	0,65	0,97
TF-IDF + LR	0,61	0,63	0,98
TF-IDF + NB	0,61	0,63	0,92
TF-IDF + SVM	0,65	0,68	0,99
TF-IDF + SGD	0,61	0,68	<u>0,99</u>
TF-IDF + RF	0,63	0,66	0,97
DistilBERT _{B, U}	0,59	0,52	0,50
DistilBERT _{B, C}	0,62	0,53	0,50
DistilBERT _{B, C, ML}	0,62	0,62	0,51
BERT _{B, U}	0,53	0,44	0,49
BERT _{B, C}	0,62	0,49	0,49
BERT _{B, U, ML}	0,61	0,51	0,50
BERT _{B, C, ML}	0,62	0,46	0,51
RoBERTa _B	0,54	0,41	0,49
DeBERTa _B	0,62	0,63	0,50
BERT _{L, U}	0,62	0,62	0,52
BERT _{L, C}	0,62	0,62	0,52
RoBERTa _L	0,62	0,62	0,52
DeBERTa _L	0,62	0,62	0,52

Cuadro 6.1: Resultados de *accuracy* para los diferentes modelos y *datasets*. Los valores subrayados corresponden a la mejor métrica obtenida para el *dataset*. Los subíndices utilizados para cada modelo M indican respectivamente M_B : BASE; M_L : LARGE; M_C : CASED; M_U : UNCASED; M_{ML} : MULTILINGUAL.

TF-IDF + NB respectivamente. Aún así, los demás modelos tienen un rendimiento algo similar aunque con peores métricas en los dos primeros *datasets*. En el caso del último *dataset* (News), observamos una clara distinción entre modelos BoW y TF-IDF vs. *transformers*, donde los primeros tienen unas métricas cercanas a 1 y los otros rondan los 0,5.

Un hecho curioso que encontramos en esta tabla es que sorprendentemente los modelos CASED (DistilBERT_{BASE, CASED}; DistilBERT_{BASE, CASED, MULTILINGUAL}; BERT_{BASE, CASED}; BERT_{BASE, CASED, MULTILINGUAL}) funcionan mejor que sus contrapartes UNCASED. Creemos que mantener las mayúsculas permite al modelo distinguir entre entidades (por ejemplo, ‘mar’ y ‘Mar’), ayudando a la interpretación.

Distinguiendo entre P-S_{One} y P-S_{All}, observamos también un decrecimiento en *precision* para los modelos BERT y RoBERTa_{BASE}, los cuales llegan a perder hasta 0,13 puntos pese a tener más información. Creemos que se puede deber a la forma en la que está tokenizada la información: al tener las evidencias concatenadas una detrás de otra, el modelo no consigue entender la relación entre evidencias, obteniendo un resultado indeseado.

Estos resultados de *precision* nos indica que hay aún trabajo por hacer, ya que el modelo no es capaz de encontrar las *features* que hacen que una noticia sea falsa.

A continuación, comentaremos los valores de *specificity*, donde podemos observar que

Model	P-S _{One}			P-S _{All}			News		
	Prec	Spec	F1	Prec	Spec	F1	Prec	Spec	F1
BoW + LR	0,69	0,79	0,73	0,72	0,72	0,72	0,99	0,98	0,98
BoW + NB	0,72	0,67	0,69	0,76	0,77	0,76	0,94	0,97	0,96
BoW + SVM	0,68	0,79	0,73	0,73	0,78	0,75	0,97	0,99	0,98
BoW + SGD	0,73	0,73	0,73	0,73	0,78	0,75	0,98	0,98	0,98
BoW + RF	0,63	0,95	0,76	0,65	0,95	0,77	0,99	0,94	0,97
TF-IDF + LR	0,62	0,96	0,76	0,63	0,97	0,77	0,99	0,96	0,97
TF-IDF + NB	0,62	0,99	0,76	0,62	1,00	0,77	0,99	0,80	0,88
TF-IDF + SVM	0,68	0,82	0,74	0,69	0,88	0,77	0,99	0,97	0,98
TF-IDF + SGD	0,66	0,79	0,72	0,71	0,82	0,76	0,99	0,98	0,98
TF-IDF + RF	0,64	0,93	0,76	0,66	0,95	0,78	0,99	0,94	0,97
DistilBERT _{B, U}	0,66	0,71	0,69	0,62	0,59	0,60	0,53	0,52	0,52
DistilBERT _{B, C}	0,63	0,96	0,76	0,64	0,55	0,59	0,52	0,52	0,52
DistilBERT _{B, C, ML}	0,62	1,00	0,77	0,62	1,00	0,77	0,53	0,53	0,53
BERT _{B, U}	0,62	0,65	0,63	0,56	0,49	0,52	0,51	0,51	0,51
BERT _{B, C}	0,62	1,00	0,77	0,58	0,63	0,61	0,51	0,51	0,51
BERT _{B, U, ML}	0,70	0,63	0,67	0,60	0,61	0,61	0,52	0,51	0,52
BERT _{B, C, ML}	0,69	0,69	0,69	0,56	0,61	0,58	0,53	0,53	0,53
RoBERTa _B	0,64	0,59	0,62	0,53	0,43	0,47	0,52	0,51	0,51
DeBERTa _B	0,62	1,00	0,77	0,65	0,88	0,75	0,52	0,51	0,51
BERT _{L, U}	0,62	1,00	0,77	0,62	1,00	0,77	0,52	1,00	0,69
BERT _{L, C}	0,62	1,00	0,77	0,62	1,00	0,77	0,52	1,00	0,69
RoBERTa _L	0,62	1,00	0,77	0,62	1,00	0,77	0,52	1,00	0,69
DeBERTa _L	0,62	1,00	0,77	0,62	1,00	0,77	0,52	1,00	0,69

Cuadro 6.2: Resultados de *precision*, *specificity* y *F1*, para los diferentes modelos y *datasets*. Los valores subrayados corresponden a la mejor métrica obtenida para el *dataset*. Los subíndices utilizados para cada modelo *M* indican respectivamente *M_B*: BASE; *M_L*: LARGE; *M_C*: CASED; *M_U*: UNCASED; *M_{ML}*: MULTILINGUAL.

los modelos LARGE, DistilBERT_{B, C, ML}; BERT_{B, C} y DeBERTa_B obtienen un resultado perfecto. Este rendimiento decrece para los modelos BoW y TF-IDF y acaba con los BASE, que obtienen las peores métricas.

Esto a grandes rasgos nos indica que estos modelos, aunque no son capaces de distinguir las *features* que caracterizan una noticia falsa, sí que consiguen encontrar estas *features* para las noticias verdaderas.

Finalmente, con respecto a *F1-Score*, el cual nos indica como de balanceada es la tasa acierto/error entre clase positiva/negativa, observamos que a grandes rasgos obtenemos un rendimiento muy similar entre todos los modelos de alrededor de 0,7 para el caso de P-S_{One}. Para P-S_{All} observamos una disminución generalizada en los modelos BASE, siendo esta disminución aún peor en los mismos modelos pero con el *dataset* News.

En el cuadro 6.3 evaluaremos el aprendizaje de los diferentes modelos tomando en cuenta las métricas de *accuracy* para los conjuntos de entrenamiento, validación y test. Como hemos mencionado en el capítulo 5, solamente disponemos de conjunto de validación para los modelos basados en *transformers*, ya que no es necesario de tal conjunto para las implementaciones con BoW o TF-IDF.

Model	P-S _{One}			P-S _{All}			News		
	Train	Eval	Test	Train	Eval	Test	Train	Eval	Test
BoW + LR	<u>1,00</u>	–	0,65	<u>1,00</u>	–	0,66	<u>1,00</u>	–	0,99
BoW + NB	0,99	–	0,63	0,98	–	<u>0,70</u>	0,98	–	0,97
BoW + SVM	<u>1,00</u>	–	0,63	1,00	–	0,68	1,00	–	0,98
BoW + SGD	<u>1,00</u>	–	<u>0,66</u>	<u>1,00</u>	–	0,68	<u>1,00</u>	–	0,99
BoW + RF	<u>1,00</u>	–	0,63	<u>1,00</u>	–	0,65	<u>1,00</u>	–	0,97
TF-IDF + LR	0,84	–	0,61	0,86	–	0,63	0,99	–	0,98
TF-IDF + NB	0,80	–	0,61	0,70	–	0,63	0,94	–	0,92
TF-IDF + SVM	<u>1,00</u>	–	0,65	<u>1,00</u>	–	0,68	1,00	–	<u>0,99</u>
TF-IDF + SGD	<u>1,00</u>	–	0,61	<u>1,00</u>	–	0,68	<u>1,00</u>	–	<u>0,99</u>
TF-IDF + RF	<u>1,00</u>	–	0,63	<u>1,00</u>	–	0,66	<u>1,00</u>	–	0,97
DilstilBERT _{B, U}	0,54	0,66	0,59	0,56	0,68	0,52	0,50	0,99	0,50
DilstilBERT _{B, C}	0,59	0,66	0,62	0,58	<u>0,70</u>	0,53	0,50	<u>1,00</u>	0,50
DilstilBERT _{B, C, ML}	0,62	0,62	0,62	0,62	0,62	0,62	0,50	1,00	0,51
BERT _{B, U}	0,54	0,72	0,53	0,54	0,67	0,44	0,51	0,99	0,49
BERT _{B, C}	0,62	0,63	0,62	0,56	<u>0,70</u>	0,49	0,50	<u>1,00</u>	0,49
BERT _{B, U, ML}	0,54	0,67	0,61	0,56	0,63	0,51	0,50	0,99	0,50
BERT _{B, C, ML}	0,52	0,67	0,62	0,55	0,65	0,46	0,50	1,00	0,51
RoBERTa _B	0,54	<u>0,75</u>	0,54	0,52	0,66	0,41	0,50	1,00	0,49
DeBERTa _B	0,62	0,62	0,62	0,57	0,66	0,63	0,50	1,00	0,50
BERT _{L, U}	0,62	0,62	0,62	0,62	0,62	0,62	0,52	0,52	0,52
BERT _{L, C}	0,62	0,62	0,62	0,62	0,62	0,62	0,52	0,52	0,52
RoBERTa _L	0,62	0,62	0,62	0,62	0,62	0,62	0,52	0,52	0,52
DeBERTa _L	0,62	0,62	0,62	0,62	0,62	0,62	0,52	0,52	0,52

Cuadro 6.3: Resultados de *accuracy* para los diferentes modelos y *datasets* en diferentes fases del aprendizaje. Los valores subrayados corresponden a la mejor métrica obtenida para el *dataset*. Los subíndices utilizados para cada modelo *M* indican respectivamente *M_B*: BASE; *M_L*: LARGE; *M_C*: CASED; *M_U*: UNCASED; *M_{ML}*: MULTILINGUAL.

Comparando los valores de *accuracy*, podemos afirmar que ninguno de los modelos basados en *transformers* ha llegado a sobreajustar con el número de épocas determinado. Por otro lado, parece que los modelos basados en BoW y TF-IDF tienden a sobreajustar más, cosa que no es deseable de cara a generalizar con muestras no vistas anteriormente.

Por último, consideramos necesario mencionar que aunque en principio los modelos basados en BoW o TF-IDF parezcan tener un rendimiento superior a algunos modelos basados en *transformers*, hay que tener en cuenta también el hecho de que estos modelos no son capaces de ‘entender’ el contexto entre palabras/*tokens*, por lo que no es una buena opción en una aplicación de estas características.

6.2. Interpretabilidad de los modelos

Para ahondar más en los resultados obtenidos, hemos decidido utilizar la librería `shap` (acrónimo proveniente de Shapley Additive Explanations) (S. M. Lundberg & Lee, 2017). Esta librería está basada en el valor de Shapley (Shapley, 1952), un método de distribución de riquezas en la teoría de juegos cooperativos.

El valor de Shapley puede definirse como una función que a partir de las contribuciones marginales de cada jugador mide su efecto al resultado general. Esto se puede aplicar a cualquier tipo de modelo de tipo ‘caja negra’ o *black box*, modelos que debido a su sofisticación o complejidad es imposible saber a ciencia cierta cómo funciona.

Aplicado a modelos basados en *transformers* o concretamente a la familia de modelos BERT, los valores de Shapley pueden utilizarse para explicar la contribución de cada a la predicción final realizada por el modelo.

Utilizando como ejemplo la siguiente frase en una tarea de clasificación de noticias como *fake news* o *no*:

“Los alunizajes del programa Apollo entre 1969 y 1972 fueron falsificados por la NASA”

podemos calcular los valores Shapley de cada *token* comparando la predicción realizada por BERT cuando ese *token* se incluye en la frase con la predicción realizada cuando ese *token* se excluye de la frase. La diferencia entre estas dos predicciones nos da una estimación de la importancia de esa palabra para predecir si es una *fake news* o *no*.

A continuación, vamos a ver diferentes ejemplos de noticias categorizadas como verdaderas y como *fake news* y evaluaremos los resultados obtenidos mediante la librería `shap` para entender como ‘razonan’ estos modelos en su clasificación. Para ello comenzaremos explicando unas nociones básicas del funcionamiento de esta librería para facilitar la comprensión mediante un ejemplo:

La figura 6.1 muestra los valores Shapley en un caso de análisis de sentimiento. Observamos dos representaciones para los valores Shapley: en la parte superior, tenemos los *tokens* representados según su valor en un eje numérico, mientras que en la parte inferior tenemos el texto con un subrayado que varía de color según hacia qué valor de clasificación tiende (positiva/negativa) y también varía de intensidad según su aporte a la clasificación final. Podemos interpretarla de la siguiente manera:

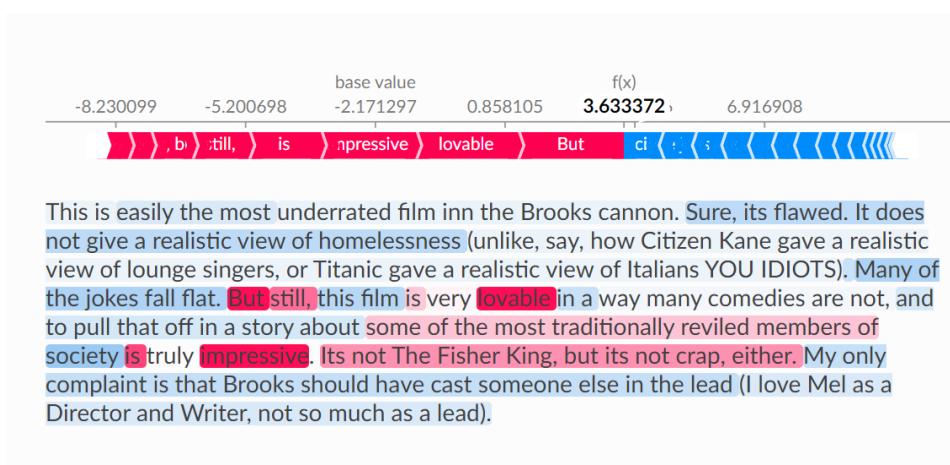


Figura 6.1: Ejemplo de visualización (S. Lundberg, 2020)

El color rojo indica tendencia a la clase positiva y el azul a la clase negativa. A partir del *base value* y $f(x)$ indicados en el eje de la parte superior, podemos conocer cuál es el resultado de la clasificación comparándolos: si $f(x) > \text{base value}$ la clasificación será positiva, mientras que si $f(x) < \text{base value}$ será negativa. Por último, podemos ver la contribución de cada token en ambas partes: en la parte superior, el tamaño de los tokens

varía según importancia o contribución a la clasificación; en la parte inferior esta relación se representa usando la intensidad de color, mayor intensidad indica mayor importancia.

Partiendo de estas nociones, podemos analizar el ‘razonamiento’ de estos modelos utilizando ejemplos de noticias del conjunto de test, mostrando algunos ejemplos representativos.

6.2.1. Politifact-Snopes One Evidence

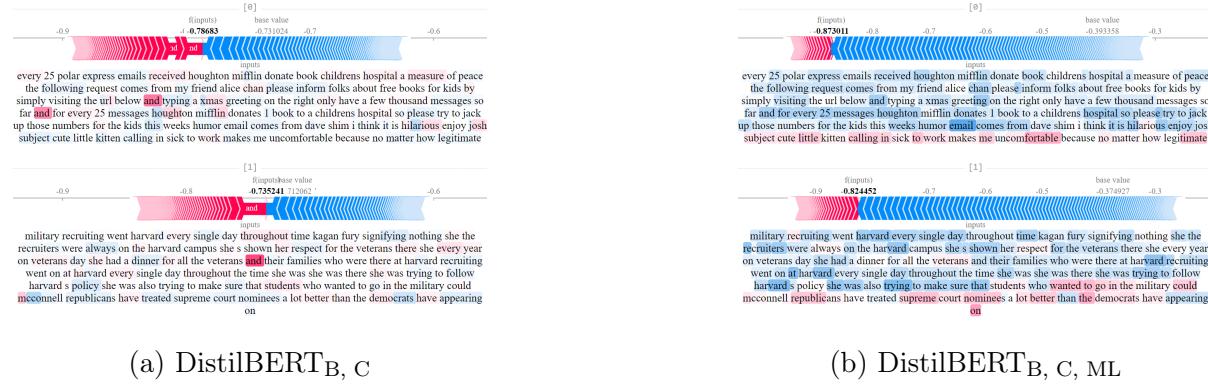


Figura 6.2: Valores Shapley de cada modelo para dos noticias: una verdadera y una falsa del dataset P-S_{One}. Los subíndices utilizados para cada modelo M indican respectivamente M_B : BASE; M_C : CASED; M_{ML} : MULTILINGUAL.

En este caso la primera noticia (superior) es verdadera, mientras que la segunda (inferior) falsa. Observamos que en ambos casos, los dos modelos tienden a la clase negativa en mayor o menor medida.

Centrándonos en los ejemplos de la figura 6.2 vemos que DistilBERT_{B, C, ML} tiene una distribución de los valores Shapley más uniforme que en DistilBERT_{B, C}. Parece que DistilBERT_{B, C, ML} ‘presta más atención’ a los *tokens* individuales y menos a los colindantes. Esto nos puede dar la intuición de que este modelo parece ‘entender’ el texto en general y por consecuencia, entender el estilo o la estructura sintáctica.

También podemos hipotetizar que este comportamiento se debe al hecho de que este dataset no tiene ningún signo de puntuación ni capitalización, por lo que en el caso de los modelos CASED no pueden hacer uso de todas estas features para clasificar las noticias.

Analizando el comportamiento de los demás modelos (véase Apéndice A.2.1.), observamos que esta tendencia es parecida para BERT_{B, C, ML}; BERT_{L, U}; BERT_{L, C}; RoBERTa_L y DeBERTa_B. Por otro lado, también encontramos otros modelos como BERT_{B, U, ML}; RoBERTa_B y DeBERTa_L que prácticamente no parecen ‘prestar atención’ a ningún token del texto. En estos casos los valores de *base value* y $f(x)$ (*logits*) también son muy similares, por lo que pensamos que con estos ejemplos los modelos no son capaces de encontrar las *features* que caracterizan o no una *fake news*.

En conclusión, observamos que algunos de los modelos parecen ser sensibles a estructuras más complejas que el *token* individual, ya que la atención o la importancia de estos *tokens* está más repartida a lo largo de la frase. Aún así, estos resultados no parecen ser concluyentes porque esa atención no está distribuida equitativamente a lo largo de la frase y tampoco el modelo consigue clasificar correctamente estos fragmentos.

6.2.2. Politifact-Snopes All Evidences

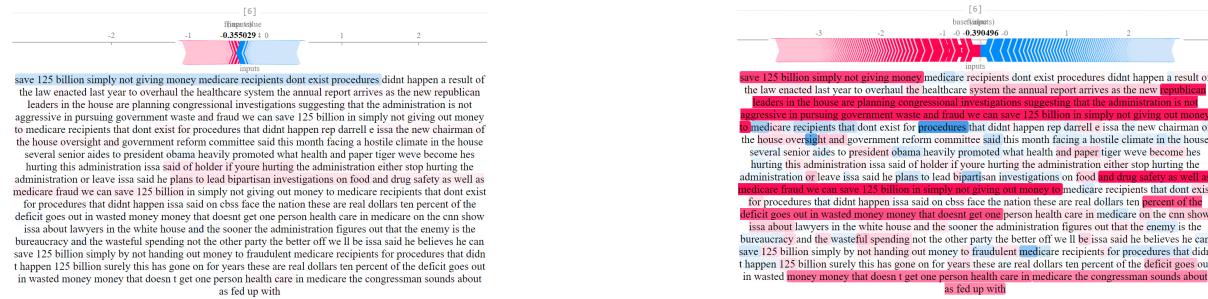


Figura 6.3: Valores Shapley de cada modelo para una noticia falsa del *dataset* P-S_{All}. Los subíndices utilizados para cada modelo *M* indican respectivamente *M_B*: BASE; *M_C*: CASED; *M_U*: UNCASED; *M_{ML}*: MULTILINGUAL.

El ejemplo elegido es de una noticia falsa o *fake news*, en la figura 6.3 notamos que BERT_{B, C, ML} parece ‘prestar atención’ a estructuras más complejas que en el caso anterior. Leyendo las partes subrayadas con más intensidad nos da a entender que el modelo está ‘entendiendo’ frases. Concretamente, estas frases son “save 125 billion simply not giving money”, “(investigations on food) and drug safety as well as medicare fraud we can save 125 billion in simply not giving out money to (medicare recipients that dont exist)”, “(ten) percent of the deficit goes out in wasted money [...].” Estas frases son muy concisas, característica clave de las *fake news*.

Por otro lado, BERT_{B, U, ML} presta muy poca atención en comparación con el ejemplo anterior, aunque de la poca atención que presta, parece hacerlo también en frases y no *tokens* individuales. Aún así, en ambos casos el resultado de la clasificación no parece influir, ya que los *base value* y *f(x)* son muy parecidos.

Los demás modelos (véase Apéndice A.2.2.) parecen seguir la tendencia de BERT_{B, C, ML}; resaltando frases concretas, estos son concretamente: BERT_{B, U}; DistilBERT_{B, C}; DistilBERT_{B, C, ML}; RoBERTa_B y DeBERTa_L. Los *logits* obtenidos, reflejados en el valor *f(x)*, generalmente son mayores a los *base values*, por lo que las clasificaciones realizadas son correctas y con un cierto grado de confianza.

El resto, sigue el comportamiento de BERT_{B, U, ML}; con poca o nula atención a frases. Los *logits* también son muy similares a los *base values*, por lo que no parece haber mucha confianza en la predicción.

6.2.3. News

Teniendo en cuenta que la noticia es verdadera, podemos observar que el peso de los *tokens* para RoBERTa_B son ínfimos en comparación con RoBERTa_L. Este último modelo no parece resaltar ningún token como relevante para hacer la predicción.

Este ejemplo no es aislado, esta tendencia es común a los siguientes modelos: BERT_{B, U}; BERT_{B, C, ML}; BERT_{L, C}; DistilBERT_{B, C}; DistilBERT_{B, C, ML}; RoBERTa_L; DeBERTa_B y DeBERTa_L. Solamente cuatro modelos de doce parecen al menos ‘razonar’ y prestan atención a tokens concretos. Aún así, este ‘razonamiento’ que tienen los otros modelos no

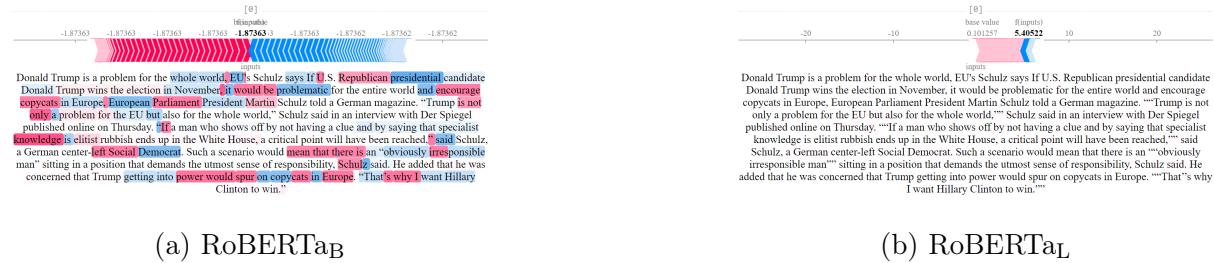


Figura 6.4: Valores Shapley para una noticia verdadera del *dataset* News. Los subíndices utilizados para cada modelo M indican respectivamente M_B : BASE; M_L : LARGE.

parece tener mucho sentido tampoco, ya que suelen prestar atención a tokens sueltos y no a estructuras sintácticas concretas, cosa que no esperábamos.

Estos modelos parecen funcionar peor con este dataset que en los casos anteriores: la atención está muy distribuida de forma muy poco uniforme, los *logits* tienen valores muy bajos en relación a los *base values* o son extremadamente mayores (esto sucede en la mayoría de modelos que ‘no prestan atención’ a ninguno de los *tokens* en el texto).

6.3. Evaluación

Habiendo analizado los resultados en la sección 6.1, podemos deducir que los modelos entrenados no parecen funcionar correctamente para detectar *fake-news* gracias a sus características estilísticas. Aún así, viendo las métricas de *specificity*, creemos que mediante esta metodología hemos creado un sistema capaz de detectar las *features* que hacen que una noticia sea verdadera, concretamente los modelos LARGE, DistilBERT_{B, C, ML}; BERT_{B, C} y DeBERTa_B son los que ofrecen mejores métricas para los *datasets* de PolitiFact-Snopes, siendo los modelos LARGE los más consistentes en todos los *datasets*.

En la sección 6.2 hemos intentando entender cuál es el funcionamiento o razonamiento interno detrás de estos modelos a la hora de clasificar. Utilizando los valores Shapley, hemos modelizado la atención de estos modelos y mediante ellos determinar qué palabras o frases son las más importantes en la clasificación.

En general, los resultados no eran los esperados, ya que las métricas obtenidas en la sección 6.1 nos daban una visión más optimista del rendimiento de estos modelos. El *dataset* con el que hemos obtenido el mejor rendimiento ha sido P-S_{All}, el siguiente P-S_{One} y el peor News.

Es por ello que creemos que es necesario investigar en profundidad otros factores resaltados en la sección 6.1, como la pérdida de rendimiento general de los modelos BASE al añadir más evidencias o al entrenar con el *dataset* News, o en la sección 6.2, como los mecanismos de atención de estos modelos y el pobre funcionamiento con el *dataset* News.

Capítulo 7

Discusión

7.1. Explicabilidad de los modelos

El artículo de Dillon et al. (2021) recalca que hay que tener cuidado al interpretar los resultados obtenidos mediante técnicas como SHAP en búsqueda de conclusiones causales, ya que estas pueden ser erróneas o basarse en hechos infundados. Debido a que los modelos predictivos asumen ciertos comportamientos iguales en el futuro, los esquemas de correlación también se mantendrán constantes.

La revisión de Mosca et al. (2022) introduce otras técnicas basadas en valores Shapley, como *L-Shapley* o *C-Shapley*, y comenta críticas de otros trabajos hacia el uso en general del uso de valores Shapley en modelos de lenguaje que debatiremos a continuación.

Kumar et al. (2020) muestra que usar los valores Shapley puede provocar inconsistencias, las cuales añaden complejidad a los modelos que intentan mitigarlas. Merrick y Taly (2019) critica la poca justificación o explicación en la incertidumbre en las explicaciones producidas mediante valores Shapley. En este caso muestran como pequeñas diferencias pueden generar efectos sobredimensionados en los resultados obtenidos, incluso en *features* que no tienen relevancia en el modelo.

Cabe recalcar que estos trabajos no están directamente aplicados a tareas de Procesado de Lenguaje Natural. Aún así, consideramos necesario tener en cuenta estas perspectivas para contextualizar sobre todas los factores que influyen en el análisis de los resultados.

7.2. Limitaciones de los modelos

El trabajo de Rogers et al. (2020) ilustra los todos los conocimientos que se saben con respecto a los modelos basados en BERT en diferentes ámbitos.

Conocimiento sintáctico. BERT no ‘entiende’ las negaciones y no es sensible al texto con una sintaxis incorrecta (orden de palabras aleatorizado, frases truncadas, sujetos o objetos eliminados (Ettinger, 2019)

Conocimiento semántico. BERT tiene dificultades a la hora de trabajar con dígitos. Esto es problema del tokenizador WordPiece, que puede dividir números de valores similares de formas muy diferentes, afectando a todos los pasos posteriores del procesamiento y finalmente al resultado.

Sentido común. BERT no consigue llevar a cabo tareas de inferencia pragmática correctamente (Ettinger, 2019) ni razonar a partir de su conocimiento del mundo (Forbes et al., 2019; Richardson & Sabharwal, 2019; W. Zhou et al., 2019)

Estos puntos mencionados dificultan el desarrollo de aplicaciones basadas en estas familias de modelos, concretamente el entrenamiento de modelos basado en hechos, ya que BERT tiene dificultades en el razonamiento. Aunque nuestra aplicación esté basada en características estilísticas, creemos que la falta de sensibilidad hacia las negaciones, estructuras sintácticas incorrectas o el problema a la hora de tratar con dígitos si que puede afectarnos en mayor o menor medida en los resultados obtenidos.

7.3. Posibles sesgos

En esta sección haremos una revisión bibliográfica sobre los trabajos realizados con respecto a determinar sesgos en LLMs.

Word Embeddings estáticos y contextuales. La investigación de Kurita et al. (2019) muestra que existen sesgos semánticos similares en *word embeddings* contextuales al igual como sucede en los *word embeddings* estáticos (Caliskan et al., 2017). Es más, el *Word Embedding Associate Test* (WEAT) (Caliskan et al., 2017), no es directamente aplicable en estos modelos, debido a su naturaleza contextual. En el caso de aplicar esta métrica para reducir los sesgos en los modelos de lenguaje contextuales no es suficiente e incluso puede agravarlos aún más (Silva et al., 2021).

Destilado de modelos. Concretamente, los modelos destilados (en el caso de DistilBERT y DistilRoBERTa) muestran un sesgo estadísticamente mayor y más fuerte que en sus versiones no destiladas (BERT y RoBERTa) (Silva et al., 2021). Esto está en línea de los hallazgos de Hooker et al. (2020), que encuentran que el proceso de destilado de modelos de visión daña desproporcionadamente a grupos minorizados. Aunque los trabajos de Gilbur (2019) y Tan y Celis (2019) concluyen que aumentar el tamaño del modelo conlleva una reducción en el sesgo, Nadeem et al. (2021) y Silva et al. (2021) demuestran lo contrario.

Tokenización. La tokenización es también crucial a la hora de desvelar sesgos: los modelos *uncased*, es decir los que no distinguen entre palabras capitalizadas y no capitalizadas suelen mostrar menos sesgos y por lo tanto mayor diversidad con respecto a nombres y pronombres (Silva et al., 2021)

Modelos aumentados. Mediante el trabajo de Vig et al. (2020), podemos vislumbrar que los LLMs, concretamente las versiones aumentadas (véase las versiones LARGE de los modelos tratados a lo largo de este trabajo) tienen mayor capacidad de ‘sintetizar’ o adquirir sesgos, aunque estos se manifiestan en una pequeña proporción de neuronas o *attention heads*. El análisis cuantitativo desempeñado muestra que pueden existir componentes en estos modelos ‘encargados’ explícitamente de reproducir sesgos o prejuicios (concretamente en este estudio solo se tratan estereotipos de género).

En conclusión, en este trabajo utilizamos modelos pre-entrenados basados en BERT, los cuales están demostrados que están sesgados. Es por ello que hay que tener en consideración para futuras investigaciones todos los factores que afectan al sesgo de los modelos para intentar reducirlo o en el caso de que no sea posible, tenerlo en cuenta en el análisis de los resultados.

Capítulo 8

Conclusiones

A continuación vamos a analizar cómo se han cumplido cada uno de los objetivos propuestos en la sección 1.2:

Definir correctamente el término *fake news* y delimitar el área de estudio a tratar. En el capítulo 2 hemos hecho una revisión bibliográfica, tomando diferentes definiciones del término y analizando los diferentes matices que aportan cada uno, descartándonos finalmente por el término de *disinformation*, trabajando solamente noticias. Sabemos que nos dejamos muchas otras formas de *disinformation*, pero es inevitable hacerlo debido a la naturaleza del Trabajo de Fin de Grado, es por ello que instamos a seguir investigando en el tema.

Desarrollar una solución que permita clasificar noticias, sirviendo como triaje en el proceso de *fact-checking*. Se han propuesto diferentes modelos de lenguaje y *datasets* y, viendo los resultados obtenidos en la sección 6.1, consideramos que hemos desarrollado un clasificador de noticias aunque los resultados no han sido los esperados. Los modelos LARGE, DistilBERT_{B, C, ML}; BERT_{B, C} y DeBERTa_B son los que mejor funcionan en este problema, aunque creemos que los modelos implementados destacan más por sus métricas de *specificity*, es decir, por su capacidad de distinguir qué características destacan en las noticias verdaderas. Aún así, por las métricas obtenidas en general, hace falta más trabajo para desarrollar una implementación que sea competente en el estado del arte actual.

Hacer un estudio comparativo de los diferentes factores de los modelos y su aprendizaje, analizando la forma en la que afectan a su rendimiento. En el capítulo 6 hemos comparado los diferentes resultados obtenidos y buscado una justificación plausible a este comportamiento. A grandes rasgos, el tamaño del modelo es el factor que parece contribuir más a la mejora de rendimiento de este. Lo segundo que parece afectar más es el hecho de si el modelo es multilingüe o distingue entre palabras capitalizadas. Por último, el tamaño del *dataset* parece ser lo que menos contribuye, aunque hay que considerar la capacidad de generalización del modelo.

Aplicar técnicas de *Explainable AI* para entender el funcionamiento interno de los modelos en el proceso de clasificación. En la sección 6.2 aplicamos estas técnicas mediante la librería `shap` y, mediante los valores Shapley, modelizamos la atención o la importancia que le da el modelo a ciertos *tokens* en la clasificación. Concluimos que el comportamiento que tienen estos modelos no es el esperado, funcionando mejor en los *datasets* P-S_{One} y P-S_{All} con menor información y número total de muestras que News.

Analizar como afectan los sesgos en esta familia de modelos a los resultados obtenidos. Como hemos observado en el apartado 7.2, gracias a una revisión bibliográfica hemos conseguido desvelar sesgos en los LLMs. Estos no solo se limitan a una representación como *word embedding* que puede dar interpretación a sesgos o prejuicios, sino que prácticamente estos sesgos son inherentes al proceso de aprendizaje. Es por ello que hay que tener en cuenta todos estos factores para realizar un análisis correcto de los resultados y de esta forma evitar perpetuando estos sesgos de forma directa e indirecta.

Capítulo 9

Trabajos futuros

Como posibles mejoras para aplicar a la metodología aplicada se sugieren las siguientes propuestas:

- Probar arquitecturas diferentes a la familia BERT, analizando su efecto en los resultados obtenidos. Un ejemplo puede ser ELECTRA (Clark et al., 2020), utilizado en el trabajo de Wang et al. (2021).
- Probar a entrenar los modelos un número mayor de épocas, implementando un mecanismo de *early stopping*.
- Realizar el experimento con *datasets* de noticias de diferentes características: mayor variedad de temas y estilos de escritura, diferencias en longitud de los textos, etc.
- Añadir más categorías o *labels* en la clasificación, como sátira.
- Aplicar otras técnicas de *Explainable AI*, como puede ser LIME.

Con respecto a nuevas líneas de investigación, se esbozan las siguientes ideas:

- Implementar *Multi-task Learning* o Aprendizaje Semi-supervisado, ya que hay indicios de ayudar en el proceso de aprendizaje y mejorar la clasificación (Rei, 2017). Wang et al. (2021) tiene una aplicación similar para clasificación de mensajes en situación de desastres.
- Generar un dataset priorizando la calidad de las noticias frente a la cantidad, similar al trabajo desarrollado por Gunasekar et al. (2023)

Bibliografía

- Abilov, A., Hua, Y., Matatov, H., Amir, O., & Naaman, M. (2021). VoterFraud2020: a Multi-modal Dataset of Election Fraud Claims on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 15, 901-912. <https://doi.org/10.1609/icwsm.v15i1.18113>
- Ahmed, H., Traore, I., & Saad, S. (2017). *Detection of misc Fake News Using N-Gram Analysis and Machine Learning Techniques* (1.^a ed., Vol. 10618). Springer, Cham. https://doi.org/10.1007/978-3-319-69155-8_9
- Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1, e9. <https://doi.org/10.1002/spy.2.9>
- Ajao, O., Bhowmik, D., & Zargari, S. (2019). Sentiment Aware Fake News Detection on Online Social Networks. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2507-2511. <https://doi.org/10.1109/ICASSP.2019.8683170>
- Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31, 211-36. <https://doi.org/10.1257/JEP.31.2.211>
- Allen, J. (1987). *Natural language understanding*. The Benjamin/Cummings Publishing Company, Inc.
- Basta, C., Costa-jussà, M. R., & Casas, N. (2019). Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 33-39. <https://doi.org/10.18653/v1/W19-3805>
- Benamira, A., Devillers, B., Lesot, E., Ray, A. K., Saadi, M., & Malliaros, F. D. (2019). Semi-Supervised Learning and Graph Neural Networks for Fake News Detection. *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 568-569. <https://doi.org/10.1145/3341161.3342958>
- Bhutani, B., Rastogi, N., Sehgal, P., & Purwar, A. (2019). Fake News Detection Using Sentiment Analysis. *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 1-5. <https://doi.org/10.1109/IC3.2019.8844880>
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.
- Blom, J. N., & Hansen, K. R. (2015). Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76, 87-100. <https://doi.org/10.1016/j.pragma.2014.11.010>
- Bommasani, R., Davis, K., & Cardie, C. (2020). Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 4758-4781. <https://doi.org/10.18653/V1/2020.ACL-MAIN.431>
- Bozarth, L., & Budak, C. (2020). Toward a Better Performance Evaluation Framework for Fake News Classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 14, 60-71. <https://doi.org/10.1609/ICWSM.V14I1.7279>

- Brady, P. (2017, septiembre). coreutils-8.28 [stable]. Consultado el 24 de junio de 2023, desde <https://lists.gnu.org/archive/html/coreutils-announce/2017-09/msg00000.html>
- Bucilă, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006*, 535-541. <https://doi.org/10.1145/1150402.1150464>
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*, 183-186. https://doi.org/10.1126/SCIENCE.AAL4230/SUPPL_FILE/CALISKAN-SM.PDF
- Castillo, C., Mendoza, M., & Poblete, B. (2011). Information Credibility on Twitter. *Proceedings of the 20th International Conference on World Wide Web*, 675-684. <https://doi.org/10.1145/1963405.1963500>
- Chandra, S., Mishra, P., Yannakoudakis, H., Nimishakavi, M., Saeidi, M., & Shutova, E. (2020). Graph-based Modeling of Online Communities for Fake News Detection. <https://arxiv.org/abs/2008.06274v4>
- Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2018). L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data. *7th International Conference on Learning Representations, ICLR 2019*. <https://arxiv.org/abs/1808.02610v1>
- Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). Misleading Online Content: Recognizing Clickbait as "False News". *Proceedings of the 2015 ACM on Workshop on Multi-modal Deception Detection*, 15-19. <https://doi.org/10.1145/2823465.2823467>
- CITS, University of California Santa Barbara. (s.f.). The Danger of Fake News in Inflaming or Suppressing Social Conflict. Consultado el 20 de junio de 2022, desde <https://www.cits.ucsb.edu/fake-news/danger-social>
- Clark, K., Luong, M.-T., Brain, G., Brain, Q. V. L. G., & Manning, C. D. (2020). ELECTRTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. <https://arxiv.org/abs/2003.10555v1>
- Cohen, S., Li, C., Yang, J., & Yu, C. (2011). Computational Journalism: A Call to Arms to Database Researchers, 148-151.
- Daniel, A., Flew, T., & Spurgeon, C. (2009). User behaviours and intentions with digital news media in Australia. *Transforming Audiences 2 ; Creativity, Knowledge, Participation*. <https://eprints.qut.edu.au/27376/>
- Davey-Attlee, F., & Soares, I. (s.f.). The fake news machine: Inside a town gearing up for 2020. Consultado el 20 de junio de 2022, desde <https://money.cnn.com/interactive/media/the-macedonia-story/>
- de Beer, D., & Matthee, M. (2021). Approaches to Identify Fake News: A Systematic Literature Review. *Integrated Science in Digital Age 2020, 136*, 13. https://doi.org/10.1007/978-3-030-49264-9_2
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1, 4171-4186. <https://arxiv.org/abs/1810.04805v2>
- Dewey, C. (2014). This is not an interview with Banksy. *The Washington Post*. Consultado el 20 de junio de 2022, desde <https://www.washingtonpost.com/news/the-intersect/wp/2014/10/21>this-is-not-an-interview-with-banksy/>
- Dillon, E., LaRiviere, J., Lundberg, S., Roth, J., & Syrgkanis, V. (2021, mayo). Be Careful When Interpreting Predictive Models in Search of Causal Insights: A Joint Article About Causality and Interpretable Machine Learning With Eleanor Dillon, Jacob LaRiviere, Scott Lundberg, Jonathan Roth, and Vasilis Syrgkanis From Microsoft.

- Consultado el 26 de junio de 2023, desde https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%5C%20careful%5C%20when%5C%20interpreting%5C%20predictive%5C%20models%5C%20in%5C%20search%5C%20of%5C%20causal%5C%C2%5C%A0insights.html
- Eisenstein, J. (2019, octubre). *Introduction to Natural Language Processing*. The MIT Press. <https://mitpress.mit.edu/9780262042840/introduction-to-natural-language-processing/>
- Elliott, V. (2022, junio). Meta Made Millions in Ads From Networks of Fake Accounts. <https://www.wired.com/story/meta-is-making-millions-from-fake-accounts/>
- Ettinger, A. (2019). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8, 34-48. https://doi.org/10.1162/tacl_a_00298
- Fallis, D. (2016). *Mis- and dis-information*. In: Floridi L (ed.) *The Routledge Handbook of Philosophy of Information (Routledge Handbooks in Philosophy)*. Routledge. <https://doi.org/10.4324/9781315757544.ch27>
- Fisher, M., Cox, J. W., & Hermann, P. (2016). Pizzagate: From rumor, to hashtag, to gunfire in D.C. *The Washington Post*. https://www.washingtonpost.com/local/pizzagate-from-rumor-to-hashtag-to-gunfire-in-dc/2016/12/06/4c7def50-bbd4-11e6-94ac-3d324840106c_story.html
- Flew, T., Spurgeon, C., Daniel, A., & Swift, A. (2012). The promise of computational journalism. *Journalism Practice*, 6(2), 157-171. <https://doi.org/10.1080/17512786.2011.616655>
- Forbes, M., Holtzman, A., & Choi, Y. (2019). Do Neural Language Representations Learn Physical Commonsense? *Proceedings of the 41st Annual Meeting of the Cognitive Science Society: Creativity + Cognition + Computation, CogSci 2019*, 1753-1759. <https://arxiv.org/abs/1908.02899v1>
- Gao, Y., Colombo, N., Holloway, R., & Wang, W. (2021). Adapting by Pruning: A Case Study on BERT. <https://arxiv.org/abs/2105.03343v1>
- Giachanou, A., Rosso, P., & Crestani, F. (2019). Leveraging Emotional Signals for Credibility Detection. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 877-880. <https://doi.org/10.1145/3331184.3331285>
- Gilbert, B. (2019, noviembre). Gender bias in GPT-2. Consultado el 25 de junio de 2023, desde <https://towardsdatascience.com/gender-bias-in-gpt-2-acf65dc84bd8>
- Giuliani-Hoffman, F. (2017). 'F*** News' should be replaced by these words, Claire Wardle says. *Cable News Network*. <https://money.cnn.com/2017/11/03/media/claire-wardle-fake-news-reliable-sources-podcast/index.html>
- Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2019). Exploring Hate Speech Detection in Multimodal Publications. *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, 1459-1467. <https://doi.org/10.1109/WACV45572.2020.9093414>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016, noviembre). *Deep Learning*. The MIT Press.
- Gottfried, J., & Shearer, E. (2016). News Use Across Social Media Platforms 2016. *Pew Research Center*. Consultado el 20 de junio de 2022, desde <https://www.pewresearch.org/journalism/2016/05/26/news-use-across-social-media-platforms-2016/>
- Gunasekar, S., Zhang, Y., Aneja, J., César, C., Mendes, T., Giorno, A. D., Gopi, S., Javaheripi, M., Kauffmann, P., De, G., Olli, R., Adil, S., Shital, S., Harkirat, S., Behl, S., Wang, X., Bubeck, S., Eldan, R., Tauman, A., ... Li, Y. (2023). Textbooks Are All You Need. <https://arxiv.org/abs/2306.11644v1>
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson,

- P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585, 357-362. <https://doi.org/10.1038/S41586-020-2649-2>
- He, P., Liu, X., Gao, J., Chen, W., & Dynamics, M. (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. <https://arxiv.org/abs/2006.03654v6>
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. <http://arxiv.org/abs/1503.02531>
- Hooker, S., Moorosi, N., Clark, G., Bengio, S., & Denton, E. (2020). Characterising Bias in Compressed Models. <https://doi.org/https://doi.org/10.48550/arXiv.2010.03058>
- Horne, B., & Adali, S. (2017). This Just In: Fake News Packs A Lot In Title, Uses Simpler, Repetitive Content in Text Body, More Similar To Satire Than Real News. *Proceedings of the International AAAI Conference on Web and Social Media*, 11, 759-766. <https://doi.org/10.1609/icwsm.v11i1.14976>
- Inc., M. P. (2023, febrero). *Meta Reports Fourth Quarter and Full Year 2022 Results*. Meta Platforms, Inc. <https://investor.fb.com/investor-news/press-release-details/2023/Meta-Reports-Fourth-Quarter-and-Full-Year-2022-Results/default.aspx>
- Ismiguzel, I. (2020, mayo). Applying Text Classification Using Logistic Regression. <https://medium.com/analytics-vidhya/applying-text-classification-using-logistic-regression-a-comparison-between-bow-and-tf-idf-1f1ed1b83640>
- Jin, Y., Wang, X., Yang, R., Sun, Y., Wang, W., Liao, H., & Xie, X. (2021). Towards Fine-Grained Reasoning for Fake News Detection. *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, 36, 5746-5754. <https://doi.org/10.1609/aaai.v36i5.20517>
- Khan, A., Brohman, K., & Addas, S. (2021). The anatomy of ‘fake news’: Studying false messages as digital objects. *Journal of Information Technology*, 37, 122-143. https://doi.org/10.1177/02683962211037693/ASSET/IMAGES/LARGE/10.1177_02683962211037693-FIG3.JPG
- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. Jupyter Notebooks – a publishing format for reproducible computational workflows (F. Loizides & B. Schmidt, Eds.). En: *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (F. Loizides & B. Schmidt, Eds.). Ed. por Loizides, F., & Schmidt, B. IOS Press. 2016, 87-90.
- Kokalj, E., Škrlj, B., Lavrač, N., Pollak, S., & Robnik-Šikonja, M. (2021). BERT meets Shapley: Extending SHAP Explanations to Transformer-based Classifiers. *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, 16-21. <https://aclanthology.org/2021.hackashop-1.3>
- Kulkarni, A., & Shivananda, A. (2019, enero). *Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python* (1.^a ed.). Apress. https://doi.org/10.1007/978-1-4842-4267-4_COVER
- Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. A. (2020). Problems with Shapley-Value-Based Explanations as Feature Importance Measures. *Proceedings of the 37th International Conference on Machine Learning*.
- Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring Bias in Contextualized Word Representations, 166-172. <https://doi.org/10.18653/v1/w19-3823>
- Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news: Addressing fake news requires a multidisciplinary effort. *Science*, 359, 1094-1096. https://doi.org/10.1126/SCIENCE.AAO2998/SUPPL_FILE/AAO2998_LAIZER_SM.PDF

- Lhoest, Q., Villanova del Moral, A., von Platen, P., Wolf, T., Šaško, M., Jernite, Y., Thakur, A., Tunstall, L., Patil, S., Drame, M., Chaumond, J., Plu, J., Davison, J., Brandeis, S., Sanh, V., Le Scao, T., Canwen Xu, K., Patry, N., Liu, S., ... Delanigue, C. (2021). Datasets: A Community Library for Natural Language Processing. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 175-184. <https://aclanthology.org/2021.emnlp-demo.21>
- Liu, Q., Yu, F., Wu, S., & Wang, L. (2017). Mining Significant Microblogs for Misinformation Identification: An Attention-based Approach. *ACM Transactions on Intelligent Systems and Technology*, 9. <https://doi.org/10.1145/3173458>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., & Allen, P. G. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. <https://arxiv.org/abs/1907.11692v1>
- Lundberg, S. (2020, diciembre). {text} plot. Consultado el 26 de junio de 2023, desde https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/text.html
- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. En I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4765-4774). Curran Associates, Inc. <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
- Ma, J., Gao, W., Joty, S., & Wong, K.-F. (2019). Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2561-2571. <https://doi.org/10.18653/v1/P19-1244>
- MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly: Management Information Systems*, 35, 293-334. <https://doi.org/10.2307/23044045>
- McCornack, S. A. (2009). Information manipulation theory. *Communication Monographs*, 59, 1-16. <https://doi.org/10.1080/03637759209376245>
- McCornack, S. A., Morrison, K., Paik, J. E., Wisner, A. M., & Zhu, X. (2014). Information Manipulation Theory 2. *Journal of Language and Social Psychology*, 33, 348-377. <https://doi.org/10.1177/0261927X14534656>
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. En S. van der Walt & J. Millman (Eds.). <https://doi.org/10.25080/Majora-92bf1922-00a>
- Merriam-Webster. (2017). How Is ‘Fake News’Defined, and When Will It Be Added to the Dictionary? Consultado el 20 de junio de 2022, desde <https://www.merriam-webster.com/words-at-play/the-real-story-of-fake-news>
- Merrick, L., & Taly, A. (2019). The Explanation Game: Explaining Machine Learning Models with Cooperative Game Theory. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, abs/1909.08128. [http://arxiv.org/abs/1909.08128](https://arxiv.org/abs/1909.08128)
- Mingers, J., & Standing, C. (2018). What is Information? Toward a Theory of Information as Objective and Veridical. *Journal on Information Technology*, 33, 85-104. <https://doi.org/10.1057/S41265-017-0038-6>
- Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022). SHAP-Based Explanation Methods: A Review for NLP Interpretability. *Proceedings of the 29th International Conference on Computational Linguistics*, 4593-4603. <https://aclanthology.org/2022.coling-1.406>

- Mourão, R. R., & Robertson, C. T. (2019). Fake News as Discursive Integration: An Analysis of Sites That Publish False, Misleading, Hyperpartisan and Sensational Information. *Journalism Studies*, 20, 2077-2095. <https://doi.org/10.1080/1461670X.2019.1566871>
- Myers, M. G., & Pineda, D. (2009). Misinformation about Vaccines. *Vaccines for Biodefense and Emerging and Neglected Diseases*, 255-270. <https://doi.org/10.1016/B978-0-12-369408-9.00017-2>
- Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5356-5371. <https://doi.org/10.18653/v1/2021.acl-long.416>
- NewsAsia, C. (2017). Government ‘seriously considering’ how to deal with fake news: Shanmugam. *Channel NewsAsia*. Consultado el 13 de junio de 2022, desde <https://web.archive.org/web/20170809191853/http://www.channelnewsasia.com/news/singapore/government--seriously-considering--how-to-deal-with-fake-news-sh-8712436?view=DEFAULT>
- Ott, M., Edunov, S., Grangier, D., & Auli, M. (2018). Scaling Neural Machine Translation. *WMT 2018 - 3rd Conference on Machine Translation, Proceedings of the Conference*, 1, 1-9. <https://doi.org/10.18653/v1/w18-6301>
- Parkinson, H. J. (2016). Click and elect: how fake news helped Donald Trump win a real election. *The Guardian*. Consultado el 20 de junio de 2022, desde <https://www.theguardian.com/commentisfree/2016/nov/14/fake-news-donald-trump-election-alt-right-social-media-tech-companies>
- Parry, J., DeMattos, E., Klementiev, A., Ind, A., Morse-Kopp, D., Clarke, G., & Palaz, D. (2022). Speech Emotion Recognition in the Wild using Multi-task and Adversarial Learning. *Interspeech 2022*, 1158-1162. <https://doi.org/10.21437/Interspeech.2022-10581>
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. En *Advances in Neural Information Processing Systems 32* (pp. 8024-8035). Curran Associates, Inc. <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12(null), 2825-2830.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1, 2227-2237. <https://doi.org/10.18653/v1/n18-1202>
- Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2016). Credibility Assessment of Textual Claims on the Web. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2173-2178. <https://doi.org/10.1145/2983323.2983661>
- Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2017). Where the truth lies: Explaining the credibility of emerging claims on the web and social media. *26th Inter-*

- national World Wide Web Conference 2017, WWW 2017 Companion*, 1003-1012. <https://doi.org/10.1145/3041021.3055133>
- Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 22-32. <https://doi.org/10.18653/V1/D18-1003>
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *Technical Report, OpenAI*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (s.f.). Language Models are Unsupervised Multitask Learners. <https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21, 1-67. <http://jmlr.org/papers/v21/20-074.html>
- Rai, A. (2017). Editor's Comments: Avoiding Type III Errors: Formulating IS Research Problems that Matter. *Management Information Systems Quarterly*, 41, v. <https://aisel.aisnet.org/misq/vol41/iss2/2>
- Read, M. (2016). Donald Trump Won Because of Facebook. *New York Magazine*. Consultado el 20 de junio de 2022, desde <https://nymag.com/intelligencer/2016/11/donald-trump-won-because-of-facebook.html>
- Rei, M. (2017). Semi-supervised Multitask Learning for Sequence Labeling. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1, 2121-2130. <https://doi.org/10.18653/V1/P17-1194>
- Reid, D. (2017). US readers confuse Facebook for a news outlet. *Consumer News and Business Channel*. Consultado el 20 de junio de 2022, desde <https://www.cnbc.com/2017/02/10/us-readers-confuse-facebook-for-a-news-outlet.html>
- Richardson, K., & Sabharwal, A. (2019). What Does My QA Model Know? Devising Controlled Probes using Expert Knowledge. *Transactions of the Association for Computational Linguistics*, 8, 572-588. https://doi.org/10.1162/tacl_a_00331
- Rochlin, N. (2017). Fake news: belief in post-truth. *Library Hi Tech*, 35, 386-392. <https://doi.org/10.1108/LHT-03-2017-0062>
- Rodríguez, A. R., Velázquez, D., Sabaté, J. G., Sabaté, C. N., Delgado, C. C., Curell, J. P., Iniesta, N. M., & Singla, L. R. (2022, febrero). Racismo digital y COVID-19: Discursos racistas y antirracistas en Twitter durante la pandemia. Consultado el 25 de junio de 2023, desde <https://www.uab.cat/doc/informeracismodigitalcovid19cast>
- Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8, 842-866. https://doi.org/10.1162/tacl_a_00349
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. <https://arxiv.org/abs/1910.01108v4>
- Shah, M. N., & Ganatra, A. (2022). A systematic literature review and existing challenges toward fake news detection models. *Social Network Analysis and Mining*, 12, 1-21. <https://doi.org/10.1007/S13278-022-00995-5/FIGURES/5>
- Shallue, C. J., Lee, J., Antognini, J., Sohl-Dickstein, J., Frostig, R., & Dahl, G. E. (2018). Measuring the Effects of Data Parallelism on Neural Network Training. *Journal of Machine Learning Research*, 20, 1-49. <https://arxiv.org/abs/1811.03600v3>

- Shapley, L. S. (1952). *A Value for N-Person Games*. RAND Corporation. <https://doi.org/10.7249/P0295>
- Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2021). Societal Biases in Language Generation: Progress and Challenges. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 4275-4293. <https://doi.org/10.18653/v1/2021.acl-long.330>
- Silva, A., Tambwekar, P., & Gombolay, M. (2021). Towards a Comprehensive Understanding and Accurate Evaluation of Societal Biases in Pre-Trained Transformers. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 2383-2389. <https://doi.org/10.18653/V1/2021.NAACL-MAIN.189>
- Silverman, C. (2016). This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook. *BuzzFeed News*. Consultado el 13 de junio de 2022, desde <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook>
- Skibinski, M. (2022, enero). Top brands are sending \$2.6 billion to misinformation websites each year. Consultado el 7 de junio de 2022, desde <https://www.newsguardtech.com/special-reports/brands-send-billions-to-misinformation-websites-newsguard-comscore-report/>
- Soll, J. (2016). The Long and Brutal History of Fake News. *POLITICO Magazine*. Consultado el 22 de junio de 2023, desde <https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535/>
- Subramanian, S. (2017). Inside the Macedonian Fake-News Complex. *WIRED*. Consultado el 13 de junio de 2022, desde <https://www.wired.com/2017/02/veles-macedonia-fake-news/>
- Suddaby, R. (2010). Editor's comments: Construct clarity in theories of management and organization. *Academy of Management Review*, 35, 346-357. <https://doi.org/10.5465/AMR.2010.51141319>
- Sundararajan, M., & Najmi, A. (2020, julio). The Many Shapley Values for Model Explanation. En H. D. III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (pp. 9269-9278, Vol. 119). PMLR. <https://proceedings.mlr.press/v119/sundararajan20b.html>
- Sydell, L. (2016). We Tracked Down A Fake-News Creator In The Suburbs. Here's What We Learned. *National Public Radio*. Consultado el 7 de junio de 2023, desde <https://www.npr.org/sections/alltechconsidered/2016/11/23/503146770/npr-finds-the-head-of-a-covert-fake-news-operation-in-the-suburbs>
- Szpakowski, M. (2017). Fake News Recognition. *GitHub repository*. Consultado el 13 de junio de 2023, desde <https://github.com/several27/FakeNewsRecognition/tree/e7822ea649a64b97e38c4f27a5f2d01d8555c7aa>
- Tan, Y. C., & Celis, L. E. (2019). Assessing Social and Intersectional Biases in Contextualized Word Representations. *Advances in Neural Information Processing Systems*, 32. <https://arxiv.org/abs/1911.01485v1>
- Tandoc, E. C., Jenkins, J., & Craft, S. (2018). Fake News as a Critical Incident in Journalism. *Journalism Practice*. <https://doi.org/10.1080/17512786.2018.1562958>
- Tandoc, E. C., Lim, Z. W., & Ling, R. (2017). Defining "Fake News". *Digital Journalism*, 6, 137-153. <https://doi.org/10.1080/21670811.2017.1360143>
- Taylor, W. L. (1953). "Cloze Procedure": A New Tool for Measuring Readability. *Journalism & Mass Communication Quarterly*, 30, 415-433. <https://doi.org/10.1177/107769905303000401>
- Tian, L., Zhang, X., & Lau, J. H. (2021). Rumour Detection via Zero-shot Cross-lingual Transfer Learning. *Lecture Notes in Computer Science (including subseries Lecture*

- Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12975 LNAI, 603-618.* https://doi.org/10.1007/978-3-030-86486-6_37
- Toraman, C., Şahinuç, F., & Yilmaz, E. (2022). Large-Scale Hate Speech Detection with Cross-Domain Transfer. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2215-2225. <https://aclanthology.org/2022.lrec-1.238>
- Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020, junio). *Practical Natural Language Processing* (1.^a ed.). O'Reilly Media, Inc.
- Van Rossum, G. (2020). *The Python Library Reference, release 3.8.2*. Python Software Foundation.
- Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems, 2017-December*, 5999-6009. <https://arxiv.org/abs/1706.03762v5>
- Venkatesan, S., Han, W., & Sharman, R. (2014). A Response Quality Model for Online Health Communities. *ICIS 2014 Proceedings*. <https://aisel.aisnet.org/icis2014/proceedings/ISHHealthcare/25>
- Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating Gender Bias in Language Models Using Causal Mediation Analysis. *Advances in Neural Information Processing Systems, 33*, 12388-12401. https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf
- Vlachos, A., & Riedel, S. (2014). Fact Checking: Task definition and dataset construction. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 18-22. <https://doi.org/10.3115/V1/W14-2508>
- Vo, N., & Lee, K. (2018). The Rise of Guardians: Fact-Checking URL Recommendation to Combat Fake News. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 275-284. <https://doi.org/10.1145/3209978.3210037>
- Vo, N., & Lee, K. (2020). Where Are the Facts? Searching for Fact-checked Information to Alleviate the Spread of Fake News. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 7717-7731. <https://doi.org/10.18653/V1/2020.EMNLP-MAIN.621>
- Volkova, S., Shaffer, K., Jang, J. Y., & Hodas, N. (2017). Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 647-653. <https://doi.org/10.18653/v1/P17-2102>
- Wang, C., Nulty, P., & Lillis, D. (2021). Transformer-based Multi-task Learning for Disaster Tweet Categorisation. <https://arxiv.org/abs/2110.08010v1>
- Wardle, C. (2017, febrero). Fake news. It's complicated. <https://firstdraftnews.org/articles/fake-news-complicated/>
- Wardle, C. (2018a). How We All Can Fight Misinformation. *Harvard Business Review*. <https://hbr.org/2018/07/how-we-all-can-fight-misinformation>
- Wardle, C. (2018b). The Need for Smarter Definitions and Practical, Timely Empirical Research on Information Disorder. *Digital Journalism*, 6, 951-963. <https://doi.org/10.1080/21670811.2018.1502047>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., ... Rush, A. M. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. <https://arxiv.org/abs/1910.03771v5>

- World, T. (2017). In Myanmar, fake news spread on Facebook stokes ethnic violence. *The World*. <https://theworld.org/stories/2017-11-01/myanmar-fake-news-spread-facebook-stokes-ethnic-violence>
- Wright, C., Gatlin, K., Acosta, D., & Taylor, C. (2022). Portrayals of the Black Lives Matter Movement in Hard and Fake News and Consumer Attitudes Toward African Americans. *Howard Journal of Communications*, 31, 19-41. <https://doi.org/https://doi.org/10.1080/10646175.2022.2065458>
- Wu, L., Rao, Y., Lan, Y., Sun, L., & Qi, Z. (2021). Unified Dual-view Cognitive Model for Interpretable Claim Verification. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 59-68. <https://doi.org/10.18653/v1/2021.acl-long.5>
- Wu, L., Rao, Y., Yang, X., Wang, W., & Nazir, A. (2021). Evidence-Aware Hierarchical Interactive Attention Networks for Explainable Claim Verification. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*.
- Xu, W., Wu, J., Liu, Q., Wu, S., & Wang, L. (2022). Evidence-aware Fake News Detection with Graph Neural Networks. *WWW 2022 - Proceedings of the ACM Web Conference 2022*, 2501-2510. <https://doi.org/10.1145/3485447.3512122>
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, 32. <https://arxiv.org/abs/1906.08237v2>
- You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., Hsieh, C.-J., & Berkeley, U. (2019). Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. <https://arxiv.org/abs/1904.00962v5>
- Yu, F., Liu, Q., Wu, S., Wang, L., & Tan, T. (2017). A Convolutional Approach for Misinformation Identification. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 3901-3907. <https://doi.org/10.24963/ijcai.2017/545>
- Zhang, M., Gable, G., & Rai, A. (2016). Toward Principles of Construct Clarity: Exploring the Usefulness of Facet Theory in Guiding Conceptualization. *Australasian Journal of Information Systems*, 20. <https://doi.org/10.3127/ajis.v20i0.1123>
- Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., & Shu, K. (2019). Mining Dual Emotion for Fake News Detection. *The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021*, 3465-3476. <https://doi.org/10.1145/3442381.3450004>
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., & Levy, O. (2023). LIMA: Less Is More for Alignment. <https://arxiv.org/abs/2305.11206v1>
- Zhou, W., Du, J., & Ren, X. (2019). Improving BERT Fine-tuning with Embedding Normalization. <https://arxiv.org/abs/1911.03918v2>

Apéndice A

Apéndice

A.1. Fragmentos originales

A.1.1. Fragmento 1

“news articles that are intentionally and verifiably false, and could mislead readers” — [Fragmento en castellano](#)

A.1.2. Fragmento 2

“fabricated information that mimics news media content in form but not in organisational process or intent” — [Fragmento en castellano](#)

A.1.3. Fragmento 3

“fake news appropriates the look and feel of real news; from how websites look; to how articles are written; to how photos include attributions. Fake news hides under a veneer of legitimacy as it takes on some form of credibility by trying to appear like real news. Furthermore, going beyond the simple appearance of a news item, through the use of news bots, fake news imitates news’ omnipresence by building a network of fake sites.” — [Fragmento en castellano](#)

A.1.4. Fragmento 4

“Information is propositional content in that it proposes that a specific state of the world exists: it is ‘what must be the case in the world for the sign to exist as and when it does’.” — [Fragmento en castellano](#)

A.1.5. Fragmento 5

“Misinformation. Propositional content of signs that misrepresents the state of the world without the intention to deceive [...]. One area where this [...] is quite common is health advice in online communities (Venkatesan et al.,

(2014), where many people spread false information unintentionally (Myers & Pineda, 2009).” — Fragmento en castellano

A.1.6. Fragmento 6

“Disinformation. Propositional content of signs that misrepresents the state of the world with the intention to deceive.” — Fragmento en castellano

A.1.7. Fragmento 7

“Malinformation. Propositional content of signs that truthfully represents the state of the world with the intention to deceive [...] It is often assumed that deception appears in the form of or as result of bald-face lying and other forms of disinformation. However, deception can equally happen in the form or as a result of subtle manipulation of information that does not necessarily misrepresent the world but is intended to deceive (McCormack, 2009; McCormack et al., 2014; Wardle, 2018b). Examples include half-truths and spin, which refer to incomplete or selective information provided with the intention to deceive (Fallis, 2016).” — Fragmento en castellano

A.2. Visualización SHAP

A.2.1. Politifact-Snopes One Evidence

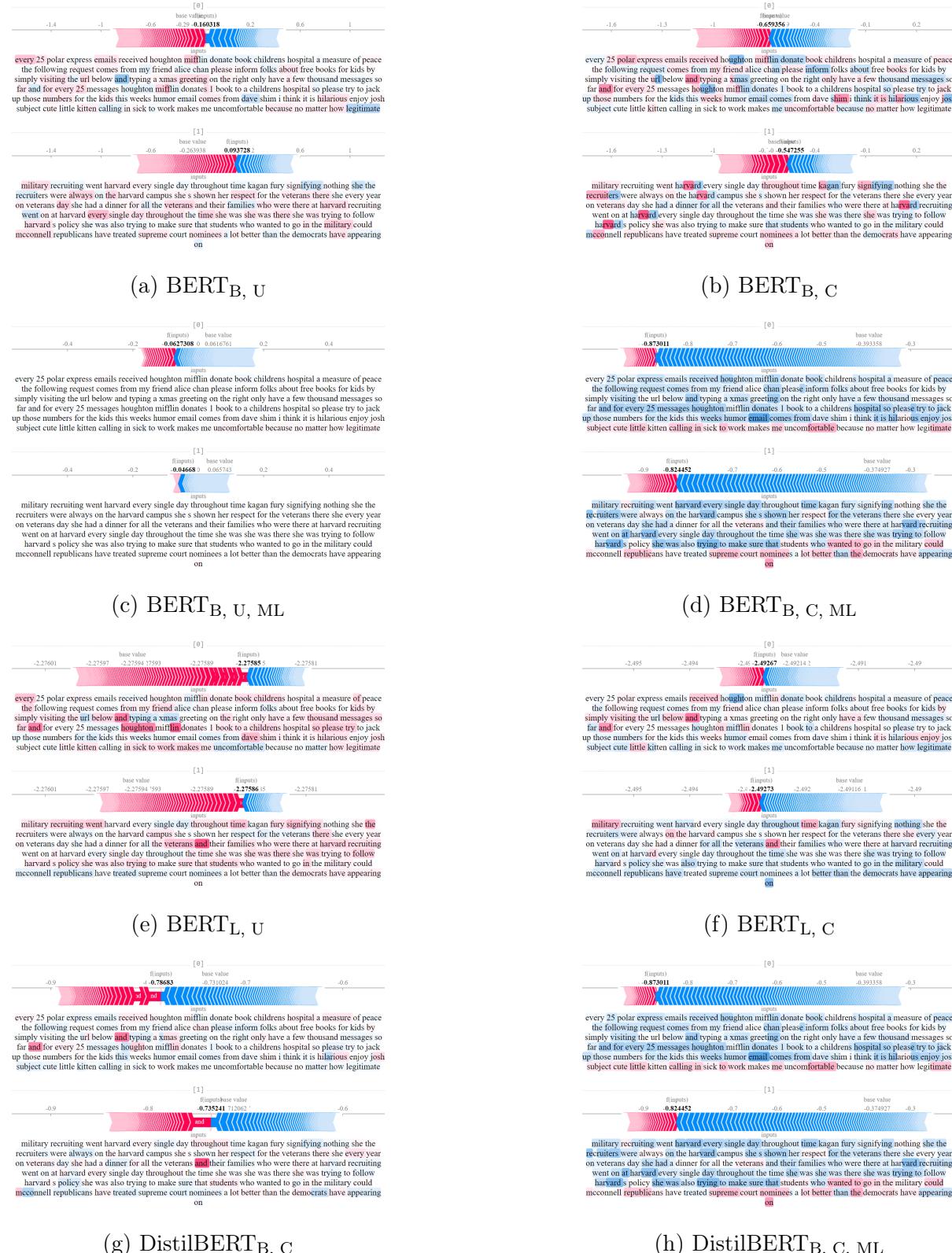




Figura A.1: Valores Shapley de cada modelo. Los subíndices utilizados para cada modelo M indican respectivamente M_B : BASE; M_L : LARGE; M_C : CASED; M_U : UNCASED; M_{ML} : MULTILINGUAL.



Figura A.2: Valores Shapley de cada modelo. Los subíndices utilizados para cada modelo M indican respectivamente M_B : BASE; M_L : LARGE; M_C : CASED; M_U : UNCASED; M_{ML} : MULTILINGUAL.

A.2.3. News

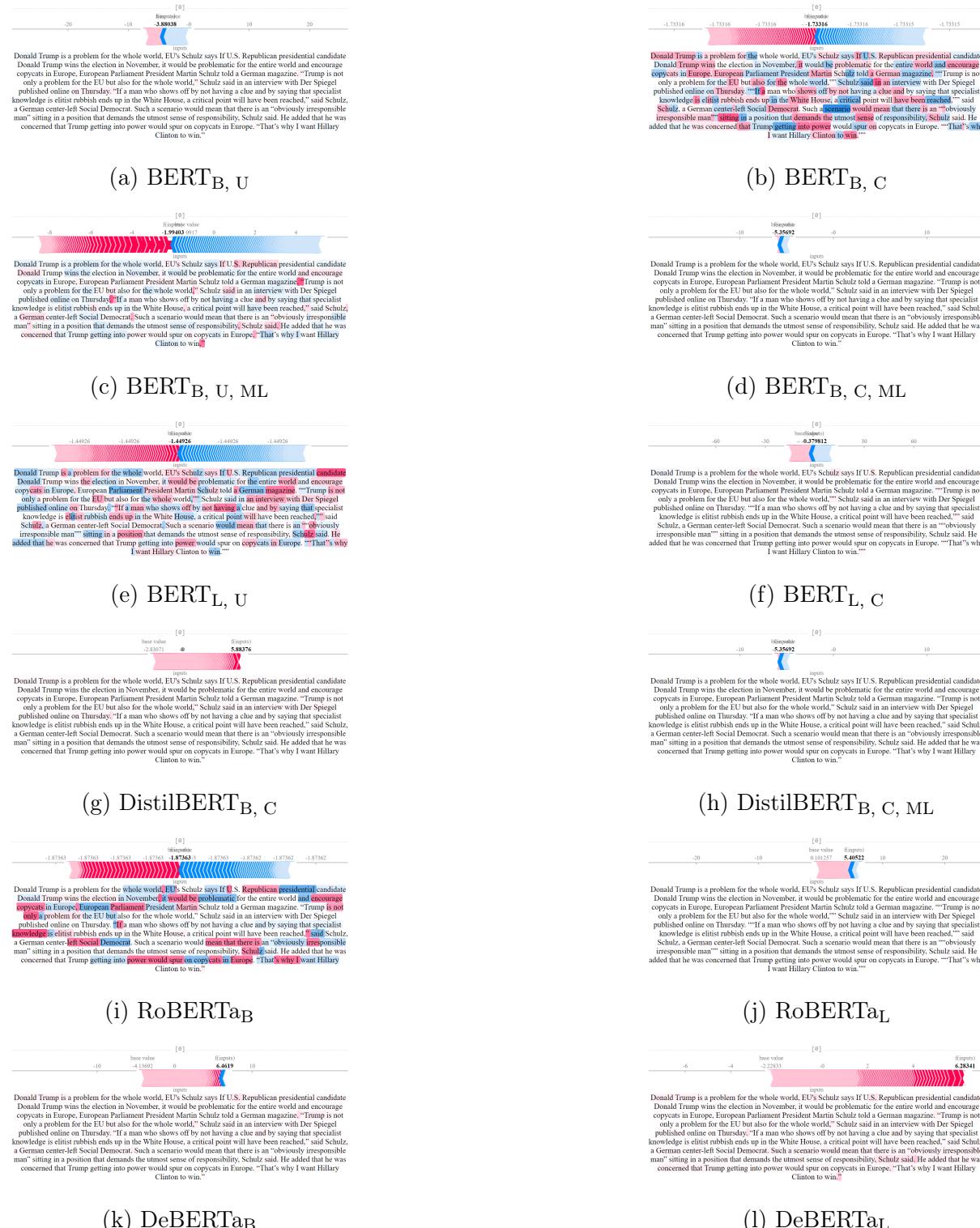


Figura A.3: Valores Shapley de cada modelo. Los subíndices utilizados para cada modelo M indican respectivamente M_B : BASE; M_L : LARGE; M_C : CASED; M_U : UNCASED; M_{ML} : MULTILINGUAL.