Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 53-62 recibido 31-03-2023 revisado 09-05-2023 aceptado 13-05-2023

Detección y clasificación de falacias prototípicas y espontáneas en español

Detection and classification of prototypical and spontaneous fallacies in Spanish

Fermín L. Cruz, José A. Troyano, Fernando Enríquez, F. Javier Ortega
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla, España
{fcruz, troyano, fenros, javierortega}@us.es

Resumen: El empleo de falacias en el seno de los debates públicos en contextos políticos, sanitarios, económicos y sociales supone un perjuicio en tanto que dificulta el entendimiento entre las partes y facilita la manipulación de la opinión pública y la propagación de desinformación. Recientemente, han aparecido conjuntos de datos que aglutinan falacias de distintos tipos, lo que habilita la experimentación en tareas como la clasificación automática de falacias. En este trabajo, presentamos el primer corpus de falacias en español, con dos secciones diferenciadas: una formada por ejemplos prototípicos extraídos de materiales educativos, y otra por ejemplos espontáneos extraídos de comentarios *on-line* a noticias. Ambas secciones incluyen ejemplos de textos no falaces, de temática similar. Los resultados preliminares al abordar las tareas de detección y clasificación usando el recurso que hemos creado muestran que se trata de una tarea desafiante (especialmente cuando se centra en falacias espontáneas) que podría ser buena candidata para formar parte de las tareas con las que se evalúan los últimos avances en modelos de lenguaje.
Palabras clave: Recursos lingüísticos, clasificación y detección de falacias, ajuste de modelos de lenguaje.

Abstract: The use of fallacies in public debates in political, health, economic and social contexts is detrimental in that it hinders understanding between the parties and facilitates the manipulation of public opinion and the propagation of misinformation. Recently, datasets containing various types of fallacies have become available, allowing experimentation in tasks such as automatic fallacy classification. In this paper, we present the first corpus of fallacies in Spanish, with two distinct sections: one formed by prototypical examples extracted from educational materials, and the other by spontaneous examples extracted from on-line comments to news items. Both sections include examples of non-fallacious texts of similar subject matter. Preliminary results on the detection and classification tasks using the corpus we have created show that it is a challenging task (especially when focused on spontaneous fallacies) that could be a good candidate to be part of the tasks with which the latest advances in language models are evaluated.
Keywords: Linguistic resources, Fallacy classification and detection, Language model tuning.

1 Introducción

Las falacias lógicas, argumentales o argumentativas (en adelante, falacias) son argumentos que parecen válidos pero que no lo son (Tindale, 2007). La invalidez de un único argumento da lugar a líneas de razonamiento erróneas, a pesar de que el resto de argumentos sean sólidos. Esto hace que la identificación de falacias sea una cuestión especialmente interesante y necesaria, ya que ayudaría a desactivar un mecanismo realmente pernicioso en el seno de los debates públicos en diversos contextos (político, sanitario, económico y social, entre otros). El uso de falacias en estos contextos, con o sin intencionalidad, suponen un claro perjuicio en el proceso comunicativo, dando lugar a situaciones indeseables como el equívoco, la desviación de la atención, el refuerzo de prejuicios y estereotipos, o incluso la manipulación del dis-

ISSN 1135-5948 DOI 10.26342/2023-71-4

curso, contaminando los debates y facilitando la propagación de noticias falsas y otros fenómenos de desinformación. Estos fenómenos tienen una gran repercusión en la sociedad y son por ello un problema de creciente interés (Allcott y Gentzkow, 2017; Vosoughi, Roy, y Aral, 2018; Kouzy et al., 2020). En ocasiones, las noticias falsas y las teorías de la conspiración se basan precisamente en el uso de falacias (Langguth et al., 2023). La disponibilidad de sistemas automáticos que permitieran detectar el uso de falacias facilitaría la moderación de contenidos en la web, y, desde un punto de vista educativo, permitiría reforzar la capacidad crítica de la población, haciéndola menos vulnerable ante desinformaciones o manipulaciones.

Este trabajo es, hasta nuestro conocimiento, el primer intento de procesar este fenómeno para textos en español. Para ello hemos recopilado y anotado un corpus de falacias, que posteriormente hemos usado para entrenar y evaluar las tareas de detección y clasificación de falacias, determinando la existencia y el tipo de entre un catálogo de tipos de falacias más comunes. Nuestro corpus contiene dos colecciones de falacias extraídas de fuentes muy distintas: las que denominamos falacias prototípicas, y las falacias espontáneas. Las falacias prototípicas han sido traducidas y posteriormente revisadas a partir del corpus LOGIC (Jin et al., 2022), un recurso compuesto por falacias en inglés extraídas de materiales educativos. Las falacias espontáneas, por su parte, han sido extraídas de conversaciones reales on-line entre usuarios. Son precisamente estas falacias espontáneas las que plantean un problema más difícil, pero a la vez más útil e interesante. En nuestros experimentos, hemos realizado un ajuste fino usando los ejemplos de nuestro corpus sobre dos modelos de lenguaje basados en la arquitectura *transformers*: Flan-T5 (Chung et al., 2022), un modelo multilingüe ajustado mediante instrucciones, y RoBERTa-BNE (Gutiérrez Fandiño et al., 2022), un modelo pre-entrenado sobre un corpus de textos exclusivamente en español.

1.1 El concepto de falacia

El concepto de falacia se conoce desde la antigüedad: ya Aristóteles hizo una primera catalogación de 13 falacias lógicas. Desde entonces, ese catálogo no ha hecho más que crecer y estructurarse. Las falacias se suelen clasificar en formales y no formales. Las falacias formales son aquellas que pueden detectarse sustituyendo las premisas por símbolos y aplicando reglas lógicas. Las no formales, por su parte, son aquellas en las que esto no ocurre. Falacias formales son, por ejemplo, las que parten de un condicional lógico o implicación ("si está nevando, entonces hace frío") y aplican una inferencia incorrecta, por ejemplo, mediante la *negación del antecedente*, infiriendo incorrectamente la *negación del consecuente* ("si no está nevando, entonces no hará frío"). En cuanto a las falacias no formales, existen multitud de clasificaciones y tipologías. Algunos tipos frecuentes son la falacia *ad hominem*, consistente en desacreditar a la persona que argumenta, y no al argumento en sí; la falacia del *hombre de paja*, consistente en generar un nuevo argumento falso exagerando o caricaturizando el argumento del oponente, para pasar a criticar este nuevo argumento falso; la falacia de *autoridad*, que vincula la veracidad de un argumento a la autoridad de quien lo esgrime (o viceversa); o la falacia *ad antiquitatem*, donde se apela a la antigüedad o la tradición como elemento verificador de un argumento.

Cabe señalar que el uso de un argumento falaz no implica la falsedad intrínseca de las conclusiones: simplemente, la argumentación no está construida de manera correcta. Esto es importante a la hora de anotar posibles falacias, pues no se debe juzgar si la conclusión es o no cierta, sino si la argumentación es sólida. Tampoco se debe considerar falaz cualquier argumentación que no constituya una prueba científica e irrefutable de lo que se defiende, siempre que quede claro en la exposición que se está enunciando una opinión o una mera posibilidad, y no una certeza sin argumentos sólidos. Todas estas consideraciones hacen que la tarea de anotación de falacias en textos y, consecuentemente, la de detección y clasificación automática, sean de una pronunciada dificultad.

1.2 Colecciones de falacias

Recientemente, se han producido distintos esfuerzos orientados a la recopilación de conjuntos de falacias, primer paso fundamental para la implementación de sistemas de detección y clasificación automática. Los trabajos de Habernal et al. (2017) y Habernal, Pauli, y Gurevych (2018) se centran en la utilización de juegos serios como mecanismo para reco-

pilar ejemplos de falacias. Es una alternativa interesante, dado que permite aprovechar el esfuerzo argumental y de razonamiento de los jugadores para anotar tareas que implican un esfuerzo de razonamiento importante, como ocurre con la detección de falacias. Otras propuestas se basan en la anotación manual de textos de distintas fuentes, en concreto foros de discusión en Reddit (Habernal et al., 2018) o textos periodísticos (Da San Martino et al., 2019). Por su parte, Sahai, Balalau, y Horincar (2021) también anotan manualmente comentarios en foros de discusión de Reddit, pero aplicando una idea interesante para obtener una lista de posibles candidatos a falacias: es relativamente habitual que los usuarios usen los nombres técnicos de las falacias para responder a otras intervenciones en la discusión. Usando esta información como etiquetado preliminar, los autores consiguen descargar de Reddit un gran volumen de textos que potencialmente pueden incluir falacias. Tras la selección manual de los ejemplos, obtienen un corpus con más de 3.000 textos etiquetados con 8 tipos de falacias. Por último, Jin et al. (2022) construyen dos conjuntos de datos diferenciados, uno a partir de ejemplos prototípicos de falacias extraídos de materiales docentes *on-line* (LOGIC), y otro a partir de la anotación manual de noticias de la web Climate Feedback (LOGICCLIMATE). Todos estos recursos están basados en el inglés, salvo (Habernal, Pauli, y Gurevych, 2018) que lo está en el alemán, no existiendo hasta nuestro conocimiento ningún recurso similar en español. Además, no aportan ejemplos de frases que no incluyan falacias, de un género similar a las falacias incluidas, lo que sería interesante de cara al entrenamiento y validación de detectores de falacias.

1.3 El conjunto de datos LOGIC

Una de las secciones del recurso que presentamos en este trabajo está basada en el conjunto de datos LOGIC (Jin et al., 2022). LOGIC está formado por ejemplos en inglés de falacias prototípicas extraídas de diversos materiales educativos *on-line* (Quizziz, study.com y ProProfs) destinados a la enseñanza o la evaluación de estudiantes acerca del entendimiento de las falacias. El recurso está formado por 2.449 ejemplos de falacias de 13 tipos.

Decimos que se trata de falacias prototípicas precisamente por la fuente de la que han sido extraídas. Se trata de ejemplos muy sintéticos y claros, en los que el tipo de falacia queda claramente representado (p.ej, "I met a tall man who loved to eat cheese. Now I believe that all tall people like cheese"). Esto las aleja del tipo de falacias que podemos encontrar en textos espontáneos, escritos en el contexto de un debate, que en muchos casos tienden a ser mucho más sutiles y complejas de identificar, como veremos más adelante.

En nuestro trabajo, no hemos utilizado directamente LOGIC, sino una versión revisada publicada por la organización Make Sense[1], en la que se han eliminado instancias erróneas o repetidas, y se ha corregido el texto o el tipo de algunos de los ejemplos. Esta versión consta de 2.226 ejemplos.

1.4 Clasificación automática de falacias

Hasta el momento, los resultados experimentales de clasificación automática de falacias son escasos, y todos para textos en inglés. Habernal, Pauli, y Gurevych (2018) aplica SVM (Cortes y Vapnik, 1995) y Bi-LSTM (Hochreiter y Schmidhuber, 1997), consiguiendo con este último los mejores resultados (0,421 de F1 con 6 clases). Habernal et al. (2018) usan Bi-LSTM y CNN para textos (Zhang y Wallace, 2017), abordando la tarea de clasificación binaria de un solo tipo de falacia, ad hominem (0,81 de accuracy). Da San Martino et al. (2019) usan la arquitectura BERT (Devlin et al., 2019) con distintas capas finales, implementando distintas granularidades de la tarea de clasificación (a nivel de documento, párrafo, oración y palabra); clasifican entre 18 clases, aunque no todas son falacias, pues el trabajo se centra en el análisis de técnicas de propaganda en noticias (0,6098 de F1 a nivel de frase). Sahai, Balalau, y Horincar (2021) aplica los mismos algoritmos de Da San Martino et al. (2019) para clasificar entre 8 tipos de falacias (0,5841 de F1 a nivel de comentarios). Por último, Jin et al. (2022) reporta resultados con diversos modelos de encoder y encoder-decoder basados en arquitecturas de tipo transformer, siendo los mejores resultados los obtenidos con Electra (Clark et al., 2020) (0,5877 de F1).

[1] https://github.com/tmakesense/logical-fallacy/tree/main/dataset-fixed

2 *La colección de datos FALLACYES*

Para construir nuestro recurso, al que hemos llamado FallacyES[2], hemos recopilado falacias en español de dos fuentes diferenciadas. Por un lado, hemos revisado y traducido las falacias prototípicas del conjunto de datos Logic en su versión corregida, y hemos añadido ejemplos similares no falaces, para habilitar la experimentación en detección además de en clasificación de falacias. Por otra parte, hemos localizado y anotado ejemplos de falacias espontáneas a partir de los comentarios a noticias publicadas en la web de agregación de noticias meneame.net[3]. También en este caso hemos incluido ejemplos de textos no falaces extraídos de los mismos comentarios. La inclusión de este tipo de falacias surge de la observación del carácter en ocasiones poco realista de las falacias que llamamos prototípicas; cabe plantearse hasta qué punto dicho tipo de falacias serían de utilidad en la detección de falacias en textos reales, asunto que abordaremos en la sección 3. Todo el proceso de anotación fue llevado a cabo por cuatro anotadores, dividiendo las tareas en grupos iguales y realizando reuniones posteriores para revisar los resultados obtenidos.

A continuación, explicamos detalladamente el procedimiento seguido para la obtención de ambas secciones del conjunto de datos.

2.1 Falacias prototípicas

La traducción de las falacias del recurso Logic ha sido llevada a cabo mediante la herramienta DeepL[4], y un proceso de revisión manual posterior de todas las traducciones para corrección de errores. A pesar de que partimos de la versión corregida del recurso, hemos encontrado nuevos errores que han sido corregidos: textos que consistían en la definición o mención explícita de la falacia en cuestión (64 instancias), textos formulados como preguntas de un cuestionario (14), textos erróneos o incompletos (5), ejemplos duplicados (4), ejemplos que no pueden considerarse falacias a falta de mayor contexto (12), y ejemplos cuya traducción no es viable (17). En este último caso, se trata de ejemplos de falacias de tipo *equivocación*, categoría que se basa en gran parte en usar la polisemia de algún término para realizar razonamientos incorrectos; estos juegos de palabras en ocasiones son imposibles de traducir. El bajo número de ejemplos finales que obtuvimos para esta categoría (27) nos hizo descartar la categoría para nuestro recurso. En total hemos eliminado 143 instancias, además de corregir el texto de 7 instancias y el tipo de falacia asignado a otras 6. Los tipos de falacias incluidos son los siguientes (se indica un acrónimo para cada uno que usaremos en las tablas y figuras en adelante):

- **generalización apresurada (ga)**: se extrae una conclusión general en base a uno o pocos casos.
- **ad hominem (ah)**: se ataca a la persona o entidad que argumenta, en lugar de razonar acerca del argumento en sí.
- **ad populum (ap)**: basa la veracidad (o falsedad) de un argumento en que la mayoría de las personas lo considera cierto (o falso).
- **falsa causalidad (fc)**: establece una relación de causalidad entre dos fenómenos sin aportar evidencias.
- **razonamiento circular (rc)**: razonamiento erróneo en varios pasos, en el que premisa y conclusión se basan una en la otra para demostrar su veracidad.
- **apelación a las emociones (ae)**: trata de manipular los sentimientos del receptor para ganar el debate.
- **pista falsa (pf)**: introduce un nuevo tema de debate no relacionado directamente con el original, distrayendo la atención sobre el debate original.
- **deducción errónea (de)**: se presenta un razonamiento en apariencia lógico, pero con errores formales que lo invalidan (falsa analogía, negación del antecedente, afirmación del consecuente,...)
- **credibilidad (c)**: basa la veracidad (o falsedad) de un argumento en la opinión de una autoridad en el tema (falacia de *autoridad*), o en la opinión tradicionalmente aceptada (falacia *ad antiquitatem*).
- **falso dilema (fd)**: se presentan dos opciones como las únicas posibles, cuando realmente existen muchas opciones posibles.
- **hombre de paja (hp)**: se reformulan los argumentos del oponente de manera exagerada o caricaturizada, para pasar a atacar esa nueva versión distorsionada de los argumentos.

[2] *https://github.com/ITALIC-US/FallacyES*
[3] http://old.meneame.net
[4] https://www.deepl.com

- **intencional (i)**: cualquier otro tipo de falacia que no entre en las categorías anteriores, pero en la que quede patente la intención del orador de ganar el debate sin usar argumentos correctos.

2.1.1 Ejemplos no falaces

Para poder realizar experimentos no sólo de clasificación sino también de detección de falacias, hemos creado manualmente ejemplos de argumentos no falaces, partiendo de las instancias anteriores. La tarea consistió en realizar transformaciones a cada uno de los enunciados, de manera que se mantuviese en la medida de lo posible la temática tratada en cada texto, pero haciendo las matizaciones o correcciones necesarias para eliminar el uso de falacias. Por ejemplo, la instancia "A Annie le debe gustar Starbucks porque a todas las chicas blancas les gusta Starbucks", que es una falacia de *generalización apresurada*, dio lugar a la instancia no falaz "A Annie le debe gustar Starbucks porque a ella le gustan las bebidas dulces y Starbucks tiene muchas opciones de bebidas dulces". Aquellas instancias en que la posible transformación resultaba demasiado complicada o se alejaba del contenido original fueron descartadas.

En la tabla 1 se muestran el número de instancias falaces de cada tipo, así como el número de instancias no falaces generadas a partir de cada tipo.

Tipo	Falacias	Longitud media	No falacias
ga	399	146±109	116
ah	258	139±92	114
ap	191	114±73	109
fc	190	155±105	116
rc	158	109±64	115
ae	140	150±123	96
pf	138	181±100	108
de	141	149±89	115
c	118	152±96	106
fd	120	92±54	109
hp	108	197±99	107
i	122	148±139	107
Total	2.083	144±100	1.318

Tabla 1: Número de falacias por tipo y de no falacias generadas a partir de cada tipo incluidas en la sección de falacias prototípicas de FallacyES, y longitud media de las falacias (en caracteres).

2.2 Falacias espontáneas

Para obtener ejemplos de falacias espontáneas (esto es, ejemplos sacados de debates reales), hemos usado como fuente los comentarios a noticias de la web meneame.net. Se trata de un agregador de noticias con un sistema de votación, en la que los usuarios pueden proponer noticias, siendo llevadas a portada aquellas noticias más votadas. La comunidad es muy propensa a comentar dichas noticias, estableciendo intensos debates alrededor de temas de actualidad política, económica, social y cultural. Basándonos en la misma idea utilizada por Sahai, Balalau, y Horincar (2021), buscamos comentarios con menciones a los nombres de los distintos tipos de falacias (comentario acusador), y almacenamos los mensajes a los que dichas menciones se refieren (comentario candidato a falacia). Por ejemplo, para el término de búsqueda "falso dilema", uno de los resultados es el siguiente:

- **Comentario acusador**: *"falacia de falso dilema. Pero no te entiendo, si le das 700 euros a alguien con ingresos eso es RBU. Yo también estoy de acuerdo en eso: para todos. para mi también."*
- **Comentario candidato a falacia**: *"Prefiero darle 700 euros a un tio que gane 1200 aunque haga alguna treta para conseguirlos que darle no se cuantos millones a OHL o a Florentino para que te construya cualquier mierda y se queden un buen pico la verdad."*

Posteriormente, utilizamos todos estos candidatos para anotar manualmente aquellos trozos de los comentarios candidatos que constituyan falacias del tipo correspondiente. Esto no ocurre en la mayoría de los casos, pues muchas veces las acusaciones de falacia no están fundamentadas, o es difícil extraer un trozo de texto con el suficiente contexto como para que la falacia sea inequívoca. Hemos sido muy rigurosos en esta selección, tratando de quedarnos sólo con ejemplos que no admitan dudas sobre la existencia de la falacia en cuestión. En total, se revisaron más de 14.000 comentarios candidatos. Aunque hemos llevado a cabo este procedimiento para los 12 tipos de falacias incluidos en la sección anterior del recurso, sólo hemos encontrado un número significativo de ejemplos para 8 de las categorías. Las categorías no incluidas son aquellas para las cuáles no hemos localizado un número suficiente de comentarios acusadores (al menos 30), o bien la mayoría de los comentarios candidatos no tenían el suficiente contexto para poder discernir ese

tipo de falacias. En la tabla 2 se muestran el número total de instancias de cada tipo de falacia obtenidas y la longitud media de las mismas.

Además del tipo y el texto, para cada una de las instancias también incluimos en el recurso el titular y el resumen de la noticia de cuyos comentarios se extrajo el texto; dicho contexto podría ser incluido como entrada en los modelos de detección y clasificación de falacias.

Tipo	**Instancias**	**Longitud media**
generalización apresurada	48	172±87
ad hominem	208	144±100
ad populum	67	127±60
apelación a las emociones	38	128±49
deducción errónea	42	224±152
credibilidad	185	164±109
falso dilema	182	128±64
hombre de paja	153	163±73
Total falacias	923	151±93
Total no falacias	917	148±63

Tabla 2: Número de instancias por tipo incluidas en la sección de falacias espontáneas de FallacyES, y longitud media de los ejemplos (en caracteres).

2.2.1 Ejemplos no falaces

Aunque inicialmente intentamos localizar ejemplos no falaces a partir de los candidatos revisados en el proceso explicado en la sección anterior, encontramos que en su mayoría tampoco son buenos candidatos para esta categoría, pues muchos bordean el uso de alguna falacia, o al menos no puede asegurarse su clase a la vista del contexto proporcionado. Para localizar ejemplos no falaces que incluir en esta sección del recurso, hemos revisado todo el hilo de comentarios de las noticias para las que hemos encontrado ejemplos de falacias, buscando trozos de participaciones que expresen opiniones o razonamientos sin caer en el uso de falacias. Al usar como fuente los comentarios de las mismas noticias, pretendemos que la temática de los ejemplos no falaces sea similar a la de los ejemplos falaces, para evitar que los modelos entrenados se basen en las posibles diferencias de terminología relacionada. También hemos analizado la distribución de longitud de las instancias de ambos tipos (falaces y no falaces), para asegurarnos de que no haya diferencias significativas entre ambas que puedan falsear los resultados de evaluación de los modelos entrenados a partir de dichos ejemplos (ver tabla 2).

3 Experimentación

En todos los experimentos, hemos procedido a realizar un ajuste fino sobre dos modelos pre-entrenados basados en *transformers*, uno de tipo *encoder-decoder* (Flan-T5) y otro de tipo *encoder* (RoBERTa-BNE), haciendo uso de los ejemplos disponibles en el conjunto de datos que hemos recopilado. Hemos elegido Flan-T5 (Chung et al., 2022) por los buenos resultados alcanzados en diversas tareas de PLN, mejorando el estado del arte en muchos casos. Los modelos Flan-T5 son versiones de los modelos originales T5 (Raffel et al., 2020) que han sido ajustados mediante *instruction fine-tuning* sobre un amplio conjunto de tareas. Como se trata de un modelo de tipo *sequence-to-sequence*, es decir, que recibe un texto de entrada y genera un texto de salida, hemos codificado las tareas de detección y clasificación de falacias de la siguiente manera:

- **Entrada**: *Posible falacia: <texto ejemplo>*
- **Salida detección**: *falacia* o *no falacia*
- **Salida clasificación**: *<tipo de falacia>*

Por su parte, hemos elegido RoBERTa-BNE (Gutiérrez Fandiño et al., 2022) por tratarse de un modelo pre-entrenado exclusivamente sobre textos en español: el corpus BNE, un enorme conjunto de textos de 540GB. Este corpus fue elaborado a partir de los textos de todas las webs con dominios .es recopilados anualmente por la Biblioteca Nacional de España desde 2009 hasta 2019. En este caso, al tratarse de un modelo de tipo *encoder*, para realizar el ajuste fino hemos añadido una capa densa final con una neurona para la tarea de detección y tantas neuronas como clases de salida para la de clasificación. De entre los modelos de distintos tamaños disponibles, hemos escogido Flan-T5-base (250 millones de parámetros) y RoBERTa-base-BNE (125 millones de parámetros), por ser los modelos de mayor tamaño sobre los que podemos ejecutar el ajuste fino con los recursos de que disponemos actualmente. Debe suponerse por tanto un margen de mejora en los resultados obtenidos, puesto que existen modelos de hasta 11 mil millones y 355 millones de parámetros

para Flan-T5 y RoBERTa-BNE, respectivamente.

Para tener una referencia basada en bolsas de palabras con la que comparar los resultados, y medir de esta forma la importancia de la estructura del discurso en las falacias, todos los experimentos han sido replicados entrenando un clasificador de tipo *Logistic Regression*, haciendo uso de pesado tf-idf y considerando unigramas y bigramas. En cada experimento, se han utilizado el 90 % de las instancias disponibles para entrenamiento y el 10 % restantes para evaluación, distribuidas aleatoriamente de manera estratificada. Se han utilizado exactamente las mismas particiones en todos los experimentos. Dado que había relativamente pocos datos disponibles, hemos optado por no definir una partición de desarrollo, usando valores por defecto para los hiperparámetros; existe por tanto en este sentido un margen de mejora con respecto a los resultados aquí mostrados. El código de toda la experimentación está disponible en GitHub para su completa reproducibilidad[5].

3.1 Clasificación

Los resultados de clasificación se muestran en la tabla 3. Hemos realizado experimentos con las falacias prototípicas (P) y con las espontáneas (E). Los números que acompañan a los nombres de los experimentos hacen referencia al número de clases consideradas, indicándose el número total de instancias en la segunda columna de la tabla. La métrica de evaluación es F1 macro-ponderada, dadas las diferencias entre el número de instancias de cada clase en algunos experimentos.

3.1.1 Flan-T5 vs RoBERTa-BNE

En todos los casos salvo en el experimento P-8 → E-8, los mejores resultados son obtenidos con un amplio margen por RoBERTa-BNE frente a Flan-T5. Es posible que el preentrenamiento sobre un gran volumen de textos exclusivamente en español sea determinante, a pesar de ser un modelo con la mitad de parámetros que Flan-T5. Por su parte, los resultados del clasificador *Logistic Regression* son muy pobres, lo que indica la imposibilidad práctica de afrontar la tarea de clasificación de falacias desde un modelo de bolsa de palabras. Esto es coherente con la importancia de la estructura discursiva inherente al concepto de falacia.

[5] *https://github.com/ITALIC-US/FallacyES*

3.1.2 Prototípicas vs Espontáneas

En relación al experimento P-12, aunque los resultados no son del todo comparables con los mostrados por Jin et al. (2022), puesto que en nuestro recurso los ejemplos han sido traducidos y en algunos casos corregidos (sólo 143 ejemplos eliminados y 11 corregidos, ver sección 2.1), la considerable mejora obtenida ($0,6775$ frente a $0,5877$) parece indicar el mejor desempeño de RoBERTa-BNE en la clasificación de falacias en español frente a Electra en la clasificación de falacias en inglés. La matriz de confusión (figura 1) muestra los resultados obtenidos por RoBERTa-BNE a nivel de cada uno de los tipos de falacias.

En el experimento P-8 hemos seleccionado únicamente las falacias de los tipos que están representados en la sección de falacias espontáneas, para que los resultados sean más fácilmente comparables con el experimento E-8. Los resultados obtenidos son claramente inferiores en todos los casos. Si bien esto podría estar relacionado con la mayor dificultad de la tarea de clasificación de falacias espontáneas frente a las prototípicas, no podemos asegurarlo, debido al diferente número de instancias en ambas secciones del conjunto de datos (1475 falacias prototípicas frente a 923 falacias espontáneas).

A la vista de la matriz de confusión del experimento E-8 (figura 2), en el experimento E-5 descartamos las falacias de los tipos que obtuvieron peores resultados (aquellos cuyo tipo predicho mayoritariamente fue incorrecto); las falacias descartadas son precisamente las que tienen un menor número de instancias (*apelación a las emociones*, *deducción errónea* y *generalización apresurada*). Los resultados obtenidos son significativamente mejores que los anteriores, tanto usando Flan-T5 como RoBERTa-BNE, lo que sugiere la necesidad de incluir más instancias de los tipos minoritarios para conseguir mejores resultados.

3.1.3 Clasificación inter-género

Para comprobar en qué medida las falacias prototípicas serían útiles para clasificar falacias espontáneas, hemos realizado un experimento inter-género, en el que utilizamos los modelos ajustados usando las falacias prototípicas y los evaluamos sobre las falacias espontáneas (P-8 → E-8). Mantenemos el mismo conjunto de evaluación utilizado en el experimento E-8. Los resultados no dejan lugar

a dudas: los modelos entrenados sobre las falacias prototípicas son de poca utilidad cuando se utilizan para clasificar ejemplos reales de falacias. De todas formas, esta conclusión debe ser considerada en el contexto de los experimentos expuestos: no es descartable que otros acercamientos puedan aprovecharse de los ejemplos prototípicos para mejorar el rendimiento de los clasificadores de falacias espontáneas.

Exper.	Inst.	LR	Flan	BNE
P-12	2.083	0,3907	0,6179	**0,6775**
P-8	1.475	0,4312	0,7259	**0,7459**
E-8	923	0,3517	0,6039	**0,6385**
E-5	795	0,3966	0,6575	**0,7575**
P-8 → E-8	1.568	0,1173	**0,3139**	0,2817

Tabla 3: Valores de F1 en la tarea de clasificación de falacias usando *Logistic Regression* (LR), Flan-T5 (Flan) y RoBERTa-BNE (BNE).

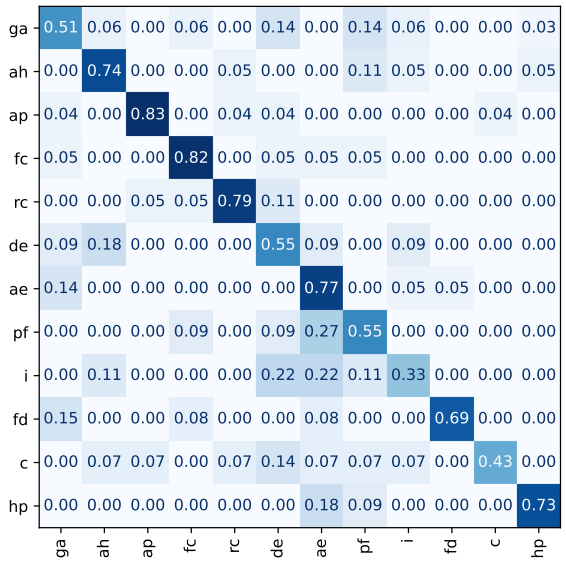

Figura 1: Matriz de confusión normalizada del experimento P-12 usando RoBERTa-BNE. Las etiquetas de los ejes horizontal y vertical hacen referencia a los tipos predichos y reales, respectivamente.

3.2 Detección

Los resultados de detección se muestran en la tabla 4. Hemos realizado tres experimentos: uno sobre las falacias prototípicas, otro con las espontáneas, y un tercero inter-género. En el caso de las falacias prototípicas, hemos utilizado un conjunto de instancias con

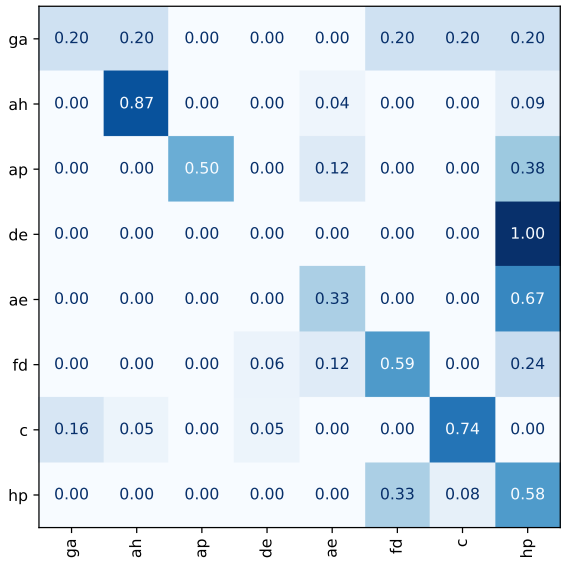

Figura 2: Matriz de confusión normalizada del experimento E-8 usando RoBERTa-BNE. Las etiquetas de los ejes horizontal y vertical hacen referencia a los tipos predichos y reales, respectivamente.

igual número de instancias de ambas clases (falacias y no falacias), evitando que se incluyan parejas de falacia y no falacia relacionadas (recordemos que las no falacias se generaron a partir de transformaciones sobre las falacias). La métrica de evaluación es F1 de la clase positiva (falacia).

Se observan mejores resultados en la detección de falacias prototípicas en comparación con las espontáneas, como ocurría con la clasificación. Sin embargo, a diferencia de dicho experimento, en este caso el número de instancias disponibles en ambas secciones del conjunto de datos es prácticamente idéntico, lo que nos permite confirmar la mayor dificultad intrínseca de las falacias espontáneas frente a las prototípicas. De nuevo, RoBERTa-BNE obtiene mejores resultados que Flan-T5, especialmente en el caso de las falacias espontáneas. En cuanto al experimento inter-género, los resultados confirman las diferencias sustanciales entre ambos tipos de falacias.

4 Conclusiones

Desde el punto de vista de su procesamiento automático, las falacias suponen un desafío semántico muy complejo dada la gran variedad de mecanismos falaces existentes, así como la sutileza que a veces requiere distinguir un argumento falaz de otro que no lo

Exper.	Inst.	LR	Flan	BNE
P	1.884	0,8137	0,9297	**0,9362**
E	1.840	0,6627	0,7892	**0,8781**
P → *E*	1.879	0,6357	**0,6698**	0,6524

Tabla 4: Valores de F1 en la tarea de detección de falacias usando *Logistic Regression* (LR), Flan-T5 (Flan) y RoBERTa-BNE (BNE).

es. La complejidad del proceso de anotación llevado a cabo para la construcción de nuestro recurso (incluyendo largas sesiones de discusión entre los anotadores para discernir la existencia y el tipo de las falacias) nos ha permitido comprobarlo empíricamente. Aún teniendo en cuenta esta dificultad intrínseca, los resultados de los experimentos realizados nos permiten ser optimistas en cuanto a las capacidades de las nuevas arquitecturas de modelos de lenguaje basados en *transformers* para detectar y clasificar falacias.

El recurso que hemos generado es, hasta nuestro conocimiento, el primer corpus de falacias en español, y el primero, independientemente de la lengua, que incluye además ejemplos de textos no falaces de la misma temática que las falacias incluidas. Esperamos que la disponibilidad pública de este recurso permita a otros investigadores avanzar en la construcción de modelos especialmente adaptados al análisis de falacias en español, permitiendo con ello facilitar la moderación de contenidos en la web y el reforzamiento de la capacidad crítica de la población, haciéndola menos vulnerable ante desinformaciones o manipulaciones.

Existe una clara correlación entre el número de instancias de cada tipo de falacia incluidos en el recurso y los resultados experimentales en las tareas de detección y clasificación, lo que nos anima a continuar aumentando el tamaño del recurso. En este sentido, nuestra intención es explorar métodos semisupervisados para encontrar nuevos ejemplos, haciendo uso de los modelos obtenidos con la versión actual del recurso. También pretendemos abordar la anotación de falacias en textos de otros géneros, con un especial interés en los debates políticos, un ámbito en el que, desgraciadamente, el uso de estos mecanismos es mucho más frecuente de lo que sería deseable.

Agradecimientos

Esta publicación es parte del proyecto PID2021-123005 financiado por MCIN/ AEI /10.13039/501100011033/ y por FEDER Una manera de hacer Europa.

Bibliografía

Allcott, H. y M. Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Chung, H. W., L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, y others. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Clark, K., M. Luong, Q. V. Le, y C. D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. En *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Cortes, C. y V. Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Da San Martino, G., S. Yu, A. Barrón-Cedeno, R. Petrov, y P. Nakov. 2019. Fine-grained analysis of propaganda in news article. En *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, páginas 5636–5646.

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, páginas 4171–4186, Minneapolis, Minnesota, Junio. Association for Computational Linguistics.

Gutiérrez Fandiño, A., J. Armengol Estapé, M. Pàmies, J. Llop Palao, J. Silveira Ocampo, C. Pio Carrino, C. Armentano Oller, C. Rodriguez Penagos, A. Gonzalez Agirre, y M. Villegas. 2022. MarIA:

Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68.

Habernal, I., R. Hannemann, C. Pollak, C. Klamm, P. Pauli, y I. Gurevych. 2017. Argotario: Computational argumentation meets serious games. En *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, páginas 7–12.

Habernal, I., P. Pauli, y I. Gurevych. 2018. Adapting serious game for fallacious argumentation to german: Pitfalls, insights, and best practices. En *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Habernal, I., H. Wachsmuth, I. Gurevych, y B. Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. En *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, páginas 386–396.

Hochreiter, S. y J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jin, Z., A. Lalwani, T. Vaidhya, X. Shen, Y. Ding, Z. Lyu, M. Sachan, R. Mihalcea, y B. Schoelkopf. 2022. Logical fallacy detection. En *Findings of the Association for Computational Linguistics: EMNLP 2022*, páginas 7180–7198, Abu Dhabi, United Arab Emirates, Diciembre. Association for Computational Linguistics.

Kouzy, R., J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, y K. Baddour. 2020. Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12(3).

Langguth, J., D. T. Schroeder, P. Filkuková, S. Brenner, J. Phillips, y K. Pogorelov. 2023. Coco: an annotated twitter dataset of covid-19 conspiracy theories. *Journal of Computational Social Science*, páginas 1–42.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, y P. J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Sahai, S. Y., O. Balalau, y R. Horincar. 2021. Breaking down the invisible wall of informal fallacies in online discussions. En *ACL-IJCNLP 2021-Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Tindale, C. W. 2007. *Fallacies and argument appraisal*. Cambridge University Press.

Vosoughi, S., D. Roy, y S. Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Zhang, Y. y B. Wallace. 2017. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. En *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, páginas 253–263, Taipei, Taiwan, Noviembre. Asian Federation of Natural Language Processing.