Uso de modelos de similitud semántica para la automatización de fact-checking: *ClaimCheck* como software de *claim matching*

Semantic similarity models for automated fact-checking: *ClaimCheck* as a claim matching tool

Irene Larraz; **Rubén Míguez**; **Francesca Sallicati**

Note: This article can be read in its English original version on:
https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/87284

Cómo citar este artículo.
Este artículo es una traducción. Por favor cite el original inglés:

Larraz, Irene; **Míguez, Rubén**; **Sallicati, Francesca** (2023). "Semantic similarity models for automated fact-checking: *ClaimCheck* as a claim matching tool". *Profesional de la información*, v. 32, n. 3, e320321.

https://doi.org/10.3145/epi.2023.may.21

Artículo recibido el 15-02-2023
Aceptación definitiva: 18-05-2023

Irene Larraz ✉
https://orcid.org/0000-0002-9164-6610

Universidad de Navarra
Campus Universitario
31009 Pamplona, España
ilarraz@alumni.unav.es

Rubén Míguez
https://orcid.org/0000-0001-9395-1849

Newtral
Vandergoten, 1
28014 Madrid, España
ruben.miguez@newtral.es

Francesca Sallicati
https://orcid.org/0000-0001-9981-8360

Newtral
Vandergoten, 1
28014 Madrid, España
francesca.sallicati@gmail.com

Resumen

Este artículo presenta el diseño experimental de *ClaimCheck*, un programa de inteligencia artificial para detectar mentiras repetidas en el discurso político a partir de un modelo de similitud semántica desarrollado por el medio de verificación *Newtral* en colaboración con *ABC Australia*. El estudio revisa el estado del arte sobre el uso de algoritmos para fact-checking y propone una definición de *claim matching*. Además, detalla el esquema de anotación de frases similares y presenta los resultados de los experimentos con el programa.

Palabras clave

Verificación; Fact-checking automatizado; *Claim matching*; Similitud semántica; Modelos de paráfrasis; Desinformación; Inteligencia artificial; IA; Algoritmos; Programas; Software.

Abstract

This article presents the experimental design of *ClaimCheck*, an artificial intelligence tool for detecting repeated falsehoods in political discourse using a semantic similarity model developed by the fact-checking organization *Newtral* in collaboration with *ABC Australia*. The study reviews the state of the art in algorithmic fact-checking and proposes a definition of claim matching. Additionally, it outlines the scheme for annotating similar sentences and presents the results of experiments conducted with the tool.

Keywords

Verification; Automated fact-checking; Claim matching; Semantic similarity; Paraphrase models; Disinformation; Artificial intelligence; AI; Algorithms; Software.

Financiación

Esta publicación es parte del proyecto de I+D+i "Medios nativos digitales en españa: tipologías, audiencias, construcción de la confianza y claves para la sostenibilidad periodística" - *Diginativemedia*. Referencia PID2021-122534OB-C22, financiado por *Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación*/10.13039/501100011033/ y *Fondo Europeo de Desarrollo Regional* "Una manera de hacer Europa".

1. Introducción

En 2020, *The Washington Post* identificó más de 50 mentiras que el expresidente de Estados Unidos Donald Trump había repetido, al menos, en una veintena de ocasiones durante su mandato. Algunas de ellas se habían pronunciado más de 200 veces (**Kessler**; **Fox**, 2021). Dos años antes, el diario había tenido que añadir una nueva categoría para catalogar las mentiras repetidas más de 20 veces. El responsable del fact-checking del medio, Glenn Kessler, explicaba que esa cifra es suficiente para demostrar que, lejos de ser un error, hay una intención de mentir, lo que termina por convertirse en propaganda (*The Washington Post*, 2018).

Los fact-checkers dedican tiempo y esfuerzos a combatir la desinformación para aportar información verificada al debate público y elevar el costo político de la mentira. Por lo que identificar las mentiras repetidas se ha convertido en una tarea prioritaria para la verificación del discurso político.

En los últimos años, los avances en el campo de la inteligencia artificial y los modelos del lenguaje a través del *deep learning* han impulsado soluciones que tratan de detectar estas repeticiones de manera automatizada. El objetivo es ayudar a los periodistas a potenciar su capacidad de monitorizar las declaraciones de los políticos, aumentar el alcance de las verificaciones y ofrecer respuestas más rápidas a las audiencias. Sin embargo, el uso de técnicas de procesamiento del lenguaje natural (o *natural language processing*, NLP) ha enfrentado distintos problemas para identificar frases similares, en especial, cuando se pronuncian con palabras o expresiones diferentes. La mayoría de los modelos que tienen este objetivo todavía no están en funcionamiento en las redacciones.

Las organizaciones de fact-checking también han comenzado a desarrollar sus propios sistemas para automatizar el *claim matching* o emparejamiento de afirmaciones, como se conoce a la tarea de identificar frases con un mismo significado. En el contexto de la verificación de datos, este concepto se refiere al proceso de comparar una afirmación con otras previamente verificadas para determinar si hay alguna coincidencia o similitud.

La organización británica de verificación *Full Fact* trabaja en un programa para revisar si una afirmación de un político se asemeja a otra que ya han verificado, contrastándola con su hemeroteca de publicaciones y alertando a los periodistas cuando detectan una repetición (**Corney**, 2021a; **Floodpage**, 2021). A la vez, el *Duke Reporters' Lab*, de la *Duke University*, está experimentando con *Squash*, una aplicación que permite encontrar afirmaciones similares entre las verificaciones de varios medios estadounidenses para lanzar una alerta en directo durante los debates electorales (**Adair**, 2021).

El propósito de este estudio es abordar los problemas de la detección de información falsa reiterada mediante el diseño de un experimento que busque validar una posible solución: la reutilización de verificaciones previas y la mejora de los sistemas de similitud semántica para realizar búsquedas de las repeticiones lingüísticas en el discurso político. La hipótesis central es que el uso de sistemas de similitud semántica puede mejorar significativamente la automatización del fact-checking, permitiendo la identificación de afirmaciones falsas que utilizan paráfrasis para expresar una misma idea. El presente artículo describe en detalle el enfoque metodológico del experimento y los resultados obtenidos, así como sus implicaciones en la investigación futura sobre la detección de información falsa en el ámbito político.

Con el objetivo de mejorar la precisión de la detección de frases similares, este estudio realiza un análisis comparativo de tres arquitecturas mediante la evaluación de su desempeño para crear *ClaimCheck*. Fueron elaboradas conjuntamente por el medio de verificación *Newtral* y la cadena *ABC Australia*, como parte del programa *JournalismAI*, coordinado por la *London School of Economics*. El sistema fue diseñado para satisfacer las necesidades específicas de los periodistas, y es el resultado de la colaboración entre ingenieros y *fact-checkers*. Asimismo, el modelo ha sido probado en el entorno de una redacción, lo que ha permitido mejorar continuamente sus resultados.

Este artículo revisa el estado del arte en la aplicación de algoritmos para la verificación (sección 2); propone una definición de *claim matching* (sección 3); detalla la metodología del diseño experimental, con un análisis de los problemas a los que se enfrentan estos modelos (sección 4), y elabora un experimento para comparar tres posibles arquitecturas (sección 5). Por último, presenta el análisis de la implementación del software en la redacción y plantea futuras vías de investigación para mejorar la tarea de *claim matching* y contribuir a la automatización del fact-checking a través de la inteligencia artificial (sección 6).

2. Estado del arte

La bibliografía académica ha documentado ampliamente la necesidad de atacar las mentiras repetidas y los problemas que genera la repetición de desinformación, como los efectos negativos de la amplificación de mensajes falsos, entre otros (**Phillips**, 2018; **Wardle**, 2018).

Babakar y **Moy** (2016) inciden en que el debate público depende en gran medida de la repetición, por lo que está presente en la hoja de ruta del fact-checking automatizado, y es uno de los problemas que aún deben resolver los investigadores. **Graves** (2018) también remarca la detección de falsedades repetidas entre los elementos fundamentales del fact-checking automatizado, y apunta a que el método más eficaz es cotejar las declaraciones con una biblioteca de frases ya verificadas por una o varias organizaciones de fact-checking.

Thorne y **Vlachos** (2018) también se refieren al *claim matching* como un tipo de modelo popular que suelen emplear las organizaciones de fact-checking cuyo proceso es el de cotejar una afirmación con otras previamente verificadas. Al igual que estos autores, **Nakov** *et al*. (2021) realizan un repaso de la tecnología para la verificación automatizada, y subrayan la detección de afirmaciones previamente verificadas como uno de los principales enfoques. Asimismo, destacan el papel del contexto para esta tarea, incluyendo el uso de frases vecinas, la resolución de correferencias y el razonamiento sobre el texto de destino.

Los efectos de las mentiras repetidas también se han analizado desde otras áreas, como la Psicología, que ha comprobado lo que se conoce como efecto de la verdad ilusoria. Este se produce cuando la repetición de un engaño genera familiaridad en quien lo escucha, reduciendo las dudas sobre su veracidad e induciendo a la idea de que lo que se oye es cierto (**Hassan**; **Barber**, 2021; **Murray** *et al*., 2020; **Agadjanian** *et al*., 2019). Algunos de estos estudios también han tratado de medir las posibilidades de que los políticos repitan una afirmación después de que se haya verificado. Por ejemplo, una investigación determinó que las probabilidades de que una afirmación se repita en los cinco días posteriores a la publicación de una verificación se reducen un 9,5% (**Lim**, 2018).

La evidencia científica también ha analizado las respuestas políticas que surgen a partir de las verificaciones, y en algunos casos se ha encontrado que los políticos continúan aferrándose a la falsedad, incluso después de haber sido desmentidos con evidencia contundente. Esto puede deberse a razones que van desde el desacuerdo hasta las estrategias basadas en la demagogia (**Sippitt**, 2020; **Porter**; **Wood**, 2021).

Por otra parte, para las ciencias computacionales, el *claim matching* se ha posicionado como uno de los principales *benchmarks* del NLP (**Nakov** *et al*., 2022). Aunque la paráfrasis y los modelos de similitud semántica todavía siguen siendo un reto por resolver, los desarrollos en los modelos del lenguaje o *large language models* (LLM) han traído consigo nuevas perspectivas para mejorar la precisión de los primeros. En concreto, el ámbito de la similitud textual semántica (STS) se centra en la tarea de medir la semejanza en el significado de las oraciones. De esta forma, los modelos examinan hasta qué punto dos afirmaciones son similares a nivel semántico, textual, léxico y referencial (**Hövelmeyer**; **Boland**; **Dietze**, 2022; **Martín** *et al*., 2021; **Sheng** *et al*., 2021).

Para entrenar estos modelos, los investigadores trabajan con *datasets* y modelos pre-entrenados como los de *SemEval-PIT2015* (en), el *STS Benchmark*, o el *Microsoft Research Paraphrase Corpus*, entre otros (**Dolan**; **Brockett**, 2005; **Lan** *et al*., 2019).

Los trabajos han tratado de resolver algunos de los problemas con los que se topa la tecnología, como la ambigüedad de las afirmaciones o la necesidad de contar con más información de contexto (**Shaar** *et al*., 2021a; 2021b; **Reimers**; **Gurevych**, 2019; **Nguyen**; **Karimi**; **Xing**, 2021). Otros han tomado aproximaciones distintas, como por ejemplo, enmarcarlo como un problema de clasificación sobre una colección de afirmaciones previamente verificadas (**Mansour**; **Elsayed**; **Al-Ali**, 2022). **Kazemi** *et al*. (2022) también proponen un modelo basado en un clasificador binario sobre *XLM-RoBERTa* (XLM-R), un popular modelo lingüístico multilingüe, y un sistema de búsqueda por similitud semántica utilizando incrustaciones de frases de modelos *LaBSE*, *SBERT* y la similitud coseno por pares.

Para ello, se han ido explorando distintos caminos, como el de *The University of Texas at Austin*, que analiza modelos estructurados y semánticos que puedan captar varios aspectos de una afirmación factual, como el tema, el tipo de dato que se expresa, las entidades implicadas y sus relaciones, cantidades, tiempos, intervalos, comparaciones y estructuras agregadas (**Arslan**, 2021). O el sistema de *Berkeley FrameNet* (**Baker**; **Fillmore**; **Lowe**, 1998), que se basa en la semántica de marcos, una vertiente de la teoría del significado, con 13 categorías de afirmaciones fácticas. Por ahora, las herramientas están diseñadas para ser sistemas híbridos en los que es necesario el juicio del periodista y la tecnología se considera un apoyo para potenciar la búsqueda y mejorar la rapidez de respuesta cuando resurge una mentira.

En el marco más general de los estudios sobre la intersección del Periodismo y la inteligencia artificial, el fact-checking ha sido objeto de un extenso análisis debido a su potencial para la automatización (*Stanford Institute for Human-Centered Artificial Intelligence*, 2023). A pesar de que aún persisten desafíos para lograr una

> Entendemos el *claim matching* como la labor de identificar afirmaciones que comparten un significado común, aunque estén expresadas de diferentes maneras

automatización completa del fact-checking, como mejorar la precisión de los programas de procesamiento del lenguaje natural y la necesidad de una base de datos de hechos verificados más completa, hay una creciente cantidad de iniciativas que hacen uso de la IA para llevar a cabo esta tarea.

En el campo del fact-checking automatizado, el *claim matching* se une a otras tareas de verificación (**Hassan** *et al.*, 2015; 2017; **Graves**, 2018; **Zeng**; **Abumansour**; **Zubiaga**, 2021) como:

- *claim detection* o detección de afirmaciones factuales dentro del discurso político;
- *check-worthiness* para medir la relevancia de una afirmación a la hora de priorizar unas verificaciones sobre otras;
- *claim validation* o contrastación del dato para confirmar la veracidad de una afirmación buscando pruebas en fuentes de datos abiertos.

Estas tareas han protagonizado la evolución de la inteligencia artificial aplicada al fact-checking, una de las áreas que más atención ha atraído gracias a sus particularidades, que permiten una mayor automatización, como la estructura de las verificaciones o la metodología que se sigue (**Nakov** *et al.*, 2021).

3. Hacia una definición de *claim matching*

Las definiciones revisadas identifican el *claim matching* como una tarea con la que se busca detectar pares de afirmaciones cuyo significado coincide aunque no lo hagan las palabras o las estructuras gramaticales que se utilizan para transmitirlo. Sin embargo, aunque existe un consenso general sobre la naturaleza de esta tarea, no hay una definición universal que delimite con precisión qué se entiende por este concepto.

Full Fact describe el *claim matching* para los propósitos de verificación como una tarea para detectar afirmaciones con una condición de verdad compartida (**Corney**, 2021b). Es decir, se consideran frases similares si ambas contienen una variable que hace que las dos sean verdaderas. Esto hace que no pueda haber una cierta y otra que no lo sea. Como ejemplo, citan la afirmación "está lloviendo en Londres", que es verdadera bajo la condición de que llueva en esa ciudad. Una frase similar puede ser "está húmedo afuera" o "está diluviando".

Por su parte, **Kazemi** *et al*. (2021) definen *claim matching* como la tarea de identificar pares de mensajes textuales que contienen afirmaciones que puedan ser atendidas con una misma verificación. Desde una perspectiva más técnica, **Shaar** *et al*. (2020) reflejan una idea similar, enmarcando la tarea como un problema de clasificación para encontrar las verificaciones que pueden ayudar a desmentir la afirmación inicial. Esto implica que la verificación contenga a la nueva afirmación. El trabajo de **Adair** *et al*. (2018) sigue esta misma línea.

La definición de **Jiang** *et al*. (2021) es más amplia e incluye

> "mensajes que contienen falsedades, inexactitudes, rumores, verdades descontextualizadas o saltos lógicos engañosos que entregan una información/tema similar con la afirmación, pero, por ejemplo, con un nombre diferente de la persona o del evento (…)".

Para estos autores, no se trata solo de identificar frases con un significado común, sino también de agrupar afirmaciones que comparten un tema o información, incluso si se presentan de manera diferente o se utilizan diferentes nombres o eventos. Esta ampliación de la definición muestra la importancia de tener en cuenta el contexto más amplio en el que se presentan, por ejemplo, cuando un político repite un mismo dato y solo cambia la referencia a la ciudad en la que está pronunciándolo.

En resumen, nuestra propuesta se fundamenta en las definiciones de los autores mencionados y busca abarcar los factores comunes en ellas. De esta forma, entendemos el *claim matching* como la labor de identificar afirmaciones que comparten un significado común, aunque estén expresadas de diferentes maneras.

4. Metodología del diseño experimental

El diseño experimental del modelo de *claim matching* se fundamentó en un proceso iterativo que involucró la evaluación y comparación de diversas arquitecturas con el fin de determinar su efectividad. Para llevar a cabo esta tarea, se realizó un análisis exhaustivo de la bibliografía científica previamente publicada en estudios relacionados y se consultó con expertos en el área de verificación de datos. A partir de este conocimiento, se procedió a la evaluación empírica de los modelos mediante la realización de pruebas utilizando un conjunto de datos de entrenamiento que contenía frases con distintos grados de similitud.

Para ello, se tomó como referencia el trabajo de **Dolan** y **Brockett** (2005) sobre el sistema *Microsoft Research Paraphrase Corpus*, así como el estudio de **Kazemi** *et al*. (2021). A partir de este corpus, elaboramos una guía de anotación en la que definimos los criterios para etiquetar los pares de afirmaciones.

"El diseño experimental del modelo de *claim matching* se fundamentó en un proceso iterativo que involucró la evaluación y comparación de diversas arquitecturas con el fin de determinar su efectividad"

Tabla 1. Ejemplos de afirmaciones anotadas y el razonamiento para la decisión

Afirmación 1	Afirmación 2	Marcado	Motivo
Más de 4.100 detenidos por violencia de género desde que comenzó el confinamiento.	Casi 9.000 detenidos por violencia de género durante el estado de alarma	Similar	Ambos mensajes contienen una cifra sobre el número de detenidos en el estado de alarma.
El Gobierno infla los datos de test realizados para presumir de que estamos en el 'top ten' mundial.	El Gobierno falsea los datos que envía a la *OCDE* para estar en el top10 de test realizados	Similar	Las dos afirmaciones apuntan a un dato alterado por el Gobierno sobre el número de tests realizados para aparecer "en el top 10".
El 20% de los contagios en España son de personal sanitario.	Por primera vez, más del 20% de la población contagiada son sanitarios.	Similar	Hay tres elementos compartidos: el porcentaje, y las referencias a los contagios y al personal sanitario.
La violencia de Género es el delito más habitual durante el estado de alarma.	La pandemia del machismo: Aumentan las agresiones a menores o personas cercanas a la mujer como forma de violencia machista, según la Fiscalía	Disímil	Ambas se refieren a violencia de género o machista durante el confinamiento, sin embargo, una apunta la frecuencia del delito y otra a dónde se produce.
282.891 parados más, y hay que sumar el millón de autónomos que han cesado su actividad y los 3,3 millones de españoles con ERTE.	Salir a presumir de gestión con más de 70.000 fallecidos, casi cuatro millones de parados, 750.000 ERTE y un millón de autónomos en el alero solo puede hacerlo quien vive en La Moncloa ajeno a los problemas de los españoles.	Similar	Aunque las dos afirmaciones aportan varios datos hay uno que coincide en ambas sobre el número de autónomos que han perdido su actividad.

Los datos de entrenamiento se obtuvieron de una base de datos de 200.000 afirmaciones factuales provistas por *Newtral*. Para cada frase se seleccionaron los tres candidatos más similares según las puntuaciones de similitud coseno, una medida utilizada para evaluar la similitud entre dos entidades en un espacio vectorial. De entre estos candidatos, se extrajeron los pares de frases para el anotado siguiendo los valores del umbral predefinido en un conjunto de datos similar (**Dolan**; **Brockett**, 2005). Según estos umbrales, seleccionamos:

- 50% de los pares con similitud de coseno >= 0,8
- 25% de los pares con similitud de coseno < 0,8 y similitud de coseno >= 0,7
- 25% de los pares con coseno similar < 0,7 y coseno similar >= 0,4

El objetivo era obtener un conjunto de datos de entrenamiento que constara de 10.000 pares de afirmaciones anotadas, incluyendo anotaciones manuales y otras con etiquetas débiles o *weak labels*, obtenidas mediante heurísticas. Estas últimas consistieron en la aplicación de reglas para identificar características comunes en las afirmaciones factuales, tales como la presencia de cifras, porcentajes, nombres geográficos y otros indicadores de factualidad. A pesar de que las etiquetas obtenidas por medio de heurísticas son menos precisas que las obtenidas manualmente, permiten ampliar significativamente el tamaño de la muestra y, por ende, mejorar la eficacia del proceso de entrenamiento.

Una vez obtenido el conjunto de datos, se procedió a la asignación de etiquetas o parámetros de similitud a las frases y a la definición de los criterios de anotación. En este proceso se consideraron diferentes enfoques tales como la asignación de valores numéricos para la medición de la similitud o la clasificación en categorías estancas para medir la similitud entre frases.

En primer lugar, se llevó a cabo una evaluación con la definición preliminar de criterios en 100 pares de afirmaciones por parte de tres anotadores diferentes que etiquetaron los pares de frases con tres categorías: similar, disímil o relacionada. También consideramos una marca N/A en caso de que la frase propuesta no fuera factual. Los resultados revelaron una alta discrepancia en las anotaciones, con un 50% de diferencias entre los anotadores. De esas frases, la mitad fue objeto de desacuerdo total entre los anotadores, ya que cada uno aplicó una etiqueta diferente. Posteriormente, los tres anotadores realizaron una revisión conjunta para tratar de alinear los criterios de anotación. Se preparó un segundo ensayo, y en esta ocasión se observó una mejora en la alineación de las anotaciones. A pesar de que persistió una discrepancia del 30%, solo un anotador marcó una etiqueta diferente en esta ocasión. Finalmente, se revisaron los criterios de anotación y se acordó eliminar una de las etiquetas ('relacionada') para evitar ruido. Con este nuevo enfoque se reanotaron los datos.

4.1. Acotar los grados de similitud: una propuesta de esquema para el anotado de datos

Acotar la definición de *claim matching* implica, a su vez, identificar qué se considera una frase similar. Existen varios factores que intervienen en este paso. El primero deriva de los problemas del uso del lenguaje, como el empleo de pronombres como "él" o referencias temporales y espaciales como "ayer" o "aquí". Un extremo de esto es las formas en las que se puede referir a una misma persona, utilizando su nombre, su apellido, las iniciales, su cargo o su anterior cargo, entre muchos otros, dificultando la identificación a los sistemas automatizados. Este problema se enmarca en otro más amplio, el de la ambigüedad de algunas afirmaciones.

> Encontrar un equilibrio entre la precisión del sistema y su capacidad para recuperar todas las frases similares es un desafío en el *claim matching*

Además de los factores mencionados, otro componente esencial para descifrar el significado de una afirmación y determinar su similitud con otras es el contexto. Este elemento es crucial en la comprensión de afirmaciones cortas en las que el contexto es difícil de extrapolar (**Shaar** *et al*., 2021b). En especial, en las afirmaciones del lenguaje oral, en las que a menudo se omiten detalles necesarios para entender a qué se está haciendo referencia. El contexto de la conversación o de quien pronuncia la afirmación puede contener información relevante que, de lo contrario, resulta difícil de deducir para los algoritmos. Por ejemplo, si lo pronuncia un político regional, probablemente haga referencia a esa región, mientras que si la misma afirmación la hace un político nacional, se extiende a ese ámbito.

Los datos en sí son otro obstáculo para recuperar frases previamente verificadas, ya que un mismo dato se puede aportar de distintas maneras. En algunos casos, los políticos pueden utilizar redondeos o cambiar de números absolutos a valores relativos, lo que puede llevar a errores de interpretación. Los verificadores han señalado que en muchos casos es más relevante la idea que los políticos buscan sustentar que el dato en sí. Un ejemplo de esto es la afirmación de algunos miembros del *Partido Popular* de España sobre el número de parados en el país, donde se utilizó una variedad de datos por encima de la cifra real. En total, diferentes integrantes del partido habían repetido al menos en 19 ocasiones que había más millones de parados de los que en realidad se contabilizan en el *Instituto Nacional de Estadística* (**Real**, 2021). Las cifras que mencionaban los representantes políticos iban desde los cuatro hasta los seis millones de parados, lo que resultaba difícil de detectar para un sistema que solo identificara la repetición de cifras.

Teniendo en cuenta lo anterior, para desarrollar *ClaimCheck* se consideró que dos frases son similares si cumplían los siguientes criterios:

- se refieren a lo mismo, aunque los datos varíen o sean incluso contradictorios;
- si una de las afirmaciones incluye detalles específicos que no invalidan a la otra;
- cuando una de las afirmaciones se refiere a una misma realidad desde una perspectiva distinta de la otra; por ejemplo, una arroja la cantidad total y la otra se refiere a la variación porcentual, pero ambas envían el mismo mensaje.

De acuerdo con los criterios establecidos, la etapa de anotación de datos utilizados para entrenar un modelo de *claim matching* es un aspecto crítico para el éxito de su implementación, dado que los parámetros y etiquetas seleccionados deben reflejar adecuadamente la similitud entre las afirmaciones. En el caso de *ClaimCheck*, se optó por una clasificación binaria con dos categorías con el objetivo de simplificar la tarea de clasificación y obtener resultados más precisos. Es importante subrayar que la selección de la categorización debe ser cuidadosamente evaluada para evitar errores de clasificación y garantizar la calidad de las anotaciones realizadas.

4.2. Problemas en la recuperación de verificaciones

Más allá de lo que se considera similar en la fase de anotación, existen problemas derivados de su aplicación específica para el fact-checking con el propósito de recuperar verificaciones previamente publicadas. En estos casos también influyen otros factores, como por ejemplo la temporalidad, ya que una frase que era falsa en un determinado momento puede ser cierta con el tiempo o al contrario. Este problema surge al recuperar verificaciones cuyos datos pueden haberse visto modificados con el paso del tiempo con estadísticas que se actualizan y valores que fluctúan, así como por las variaciones propias del contexto en el que se utilicen los datos. Es decir, la frase A y la frase B son similares, y la similitud entre ambas se mantiene en el tiempo, pero es posible que no se pueda utilizar la frase A para recuperar una verificación con la frase B si esa verificación "ha caducado", aunque eso no afecte al propio concepto o definición de similitud entre ambas.

Los matices son otro elemento determinante a la hora de identificar como similares dos frases en el campo del fact-checking, ya que la introducción de una sola palabra puede cambiar por completo el significado de la frase, alterando el sentido de la calificación o *rating* que se le ha dado antes, al pasar de ser verdadero a falso u otras categorías. Por ejemplo, la afirmación "el 50% de los nuevos contratos son indefinidos" era verdadera en cierto momento si se refería a los que se han producido desde determinada fecha, en este caso, a raíz de la reforma laboral. Sin embargo, si se elimina la palabra "nuevos", y se analiza la totalidad de los contratos, resultaba falso afirmar que "el 50% de los contratos son indefinidos". De nuevo, en este caso ambas frases son similares según los criterios establecidos en el diseño experimental, pero generaría problemas recuperar una verificación para juzgar la otra.

Sucede lo mismo con el contexto para el caso preciso de la verificación. Recuperando el ejemplo anterior, no es lo mismo que la frase "el 50% de los nuevos contratos son indefinidos" la diga un ministro, que hace referencia al ámbito nacional, que otra frase indicando que "el 50% de los contratos que hemos hecho son indefinidos" dicha por un presidente de una autonomía, que por lo general se refiere al ámbito regional de su comunidad. Ambas frases son semánticamente similares, pero el contexto hace entender que hablan de distintos datos relativos a distintos lugares. Este tipo de problemas generan desafíos a la hora de interpretar los resultados que arroja el algoritmo.

5. Método y experimentos sobre *ClaimCheck*

ClaimCheck está construido sobre la base de *ClaimHunter*, un programa de *claim detection* que identifica frases factuales en mensajes de *Twitter* (**Beltrán**; **Míguez**; **Larraz**, 2019). Se trata de algoritmos diferentes que aplican a distintas fases del proceso; mientras que el último se centra en la fase de detección de afirmaciones a verificar, el primero trabaja una vez ya está hecho ese filtro y busca casos similares ya verificados en el pasado. De esta forma, para procesar cada afirmación, el discurso político pasa primero por el algoritmo de *ClaimHunter* para corroborar si es factual y luego por el de *ClaimCheck*, que se encarga de recuperar frases candidatas a ser etiquetadas como similares.

El primer paso es identificar qué se utiliza como hemeroteca, si es solo el conjunto de datos de las verificaciones previas o también otras afirmaciones de otros políticos en redes sociales o en medios. *ClaimCheck* utiliza ambas para responder a dos objetivos:

- poder reutilizar una verificación ya hecha para actuar con más rapidez;
- identificar campañas de desinformación cuando se detectan varias afirmaciones articuladas de distintas maneras para transmitir un mismo mensaje falso.

Para ello, recopila las verificaciones del *Fact Check Explorer*, una aplicación de *Google* que almacena las publicaciones de fact-checkers que utilizan el marcado de *ClaimReview*. Este último es un sistema de datos estructurados para facilitar la lectura de las categorías que contiene una verificación, como el *claim* o afirmación que se verifica, el autor de la frase y el *rating* o valoración del grado de falsedad. Además, *ClaimCheck* utiliza también las verificaciones de otros *fact-checkers* no incluidos en la herramienta de *Google*. Como resultado, cuenta con una base de datos de 300.000 verificaciones en más de 20 idiomas, lo que en un futuro podría ampliar su alcance más allá de la verificación del discurso político, hacia la verificación de desinformación que circula en redes sociales.

A continuación, se detallan cada una de las fases: la construcción de los datasets de entrenamiento y test, el diseño de las arquitecturas y el análisis de los resultados del experimento.

5.1. Construcción de los datasets de entrenamiento y test

Para entrenar el modelo, *ClaimCheck* requirió identificar candidatos válidos, es decir, detectar frases en la base de datos de verificaciones que tuvieran el potencial de significar lo mismo que la frase de entrada. En este proceso surgen los problemas de similitud semántica mencionados previamente (ver sección 4.2), dado que se buscan frases que se asemejen. Además, el sistema de *ClaimCheck* necesitó evaluar si las dos frases se refieren exactamente a lo mismo, lo cual implica un problema de significado semántico.

El sistema de *ClaimCheck*, por lo tanto, se compone de dos partes:

- un primer componente que realiza una búsqueda de frases similares;
- un clasificador que realiza una evaluación de similitud entre el *claim* original y los candidatos propuestos.

Para construir el conjunto de datos de entrenamiento se emplearon anotaciones del *Fact Check Explorer* y, por otro lado, se seleccionaron pares del conjunto de datos de *ClaimHunter* para construir el conjunto de datos de test (*test benchmark*) con el cual poder evaluar el modelo de clasificación.

Con ello, se optó por dos estrategias. Por un lado, frases extraídas de tweets en los que podía haber más de una oración, y tweets completos en los que se incluía más información. Esta extracción de frases se hizo aplicando funciones de *tokenización* para obtener las distintas frases que componen cada tweet, y se pasaron por *ClaimHunter* para descartar frases no factuales. En concreto, a partir de la base de datos de *ClaimHunter* se generaron tres tipos de pares:

- pares de tweet con tweet;
- pares de frase con tweet;
- pares de frase con frase.

A todas estas posibles combinaciones de pares de frases se aplicaron filtros de similitud del coseno utilizando la librería *sentence-transformer* y el modelo *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2*. En concreto, se descartaron los pares con similitud del coseno menor que 0,7 y se seleccionó un subconjunto para anotación. En total, el conjunto de datos de entrenamiento constó de 7.762 pares, de los cuales un 66% eran afirmaciones similares, mientras que el test (*ClaimHunter*) *benchmark* estaba formado por 2.246 pares con un 45% de similitud. Además, se utilizó el software *Prodigy* para el anotado.

Tabla 2. Conjuntos de datos utilizados para el desarrollo de *ClaimCheck*

Dataset	Origen	Anotación	Pares similares	Pares anotados manualmente	Pares totales
Entrenamiento	*Fact Check Explorer* y otros	Similar	66%	46% *	7.760
Test	*ClaimHunter Benchmark*	Similar	45%	100%	2.246
	Prodigy Benchmark	Similar	31%	100%	7.443

*El 54% restante son *weak labels*.

5.2. Diseño de los experimentos

La experimentación con el diseño de las arquitecturas tuvo lugar entre julio y octubre de 2022 por parte de un equipo de tres investigadores que llevaron a cabo las propuestas de prototipado. Para ello, se elaboró un protocolo para la recogida de información y se diseñaron tres posibles escenarios de experimentación con sus respectivos prototipos. Cada una de las arquitecturas que se presentan a continuación supone un experimento diferente, ya que abordan los pasos de la experimentación de una manera distinta.

5.2.1. Clasificador

El primer enfoque consistió en recuperar la información con métodos estándar de *ElasticSearch*, un sistema que utiliza palabras clave y reglas para buscar los *K* candidatos más adecuados y, a continuación, filtrar utilizando un clasificador. En concreto, se trata de un clasificador tipo *BERT* para detectar pares de frases similares.

5.2.2. Búsqueda semántica + umbral

La segunda estrategia también se basó en la recuperación de información, pero esta vez mediante un enfoque de búsqueda semántica con *KNN* implementado con *OpenSearch*. El *K-Nearest Neighbors* (*KNN*) es un tipo de algoritmo de aprendizaje supervisado que se utiliza tanto para la regresión como para la clasificación. La búsqueda regular con *ElasticSearch* para encontrar coincidencias resulta útil para cuestiones generales, pero la búsqueda basada en *KNN* resulta más natural para búsquedas concretas. Esto se debe a que en esta se extraen las características del lenguaje conocidas como *embeddings* o 'representaciones vectoriales'. Este término se utiliza en el procesamiento del lenguaje natural para referirse a la técnica de representar palabras o frases como vectores numéricos, lo que permite medir la cercanía entre ellas. La similitud se establece conforme más próximos están estos valores en el espacio vectorial.

Mukherjee, **Sela** y **Al-Saadoon** (2020) citan el siguiente ejemplo: cuando buscas un vestido de novia utilizando la aplicación de búsqueda basada en *KNN*, arroja resultados similares si escribes "vestido de novia" o "vestido de matrimonio". De esta forma, "vestido de verano" y "vestido de verano floreado" son similares por la proximidad entre los *embeddings*, a diferencia de "vestido de verano" y "vestido de novia".

La forma de recuperar las frases candidatas con mayor similitud o más cercanas vectorialmente es establecer un umbral de similitud coseno para eliminar las candidatas irrelevantes. Este concepto se utiliza para establecer un valor crítico o punto de corte a partir del cual se toma una decisión o se realiza una clasificación. El problema es cómo definir un umbral de similitud adecuado para nuestra solución. En este caso, en vez de un clasificador basado en inteligencia artificial, como en el punto 1, lo que hicimos fue utilizar como clasificador simplemente el umbral de similitud. Con esto, se considera que el modelo de recuperación es lo bastante bueno, por lo que todo aquello que supere un *threshold* o umbral sería clasificado como una frase similar.

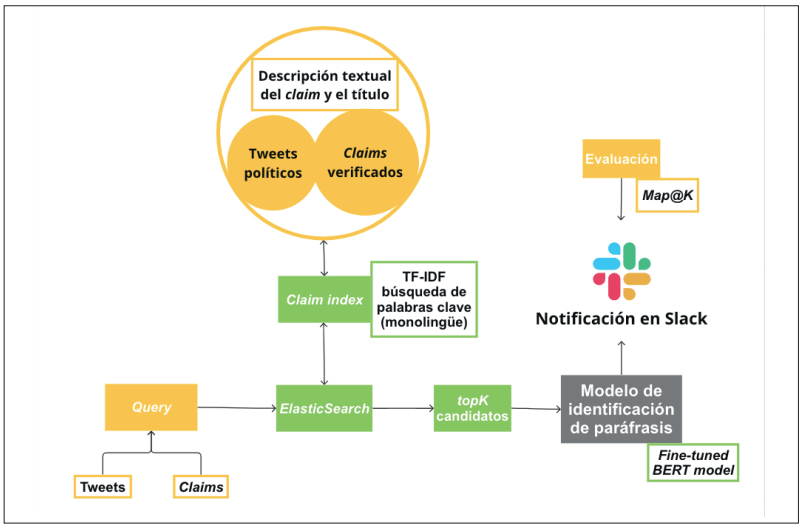

Figura 1. Arquitectura del modelo con un clasificador

5.2.3. Búsqueda semántica + clasificador

La tercera propuesta consistió en combinar dos modelos de inteligencia artificial:

- uno para generar las representaciones vectoriales en el proceso de búsqueda semántica pre-entrenado para recuperar candidatos;
- otro modelo que actúa como clasificador binario para identificar la paráfrasis entre los candidatos recuperados; es decir, para ver si las dos frases realmente dicen lo mismo o no.

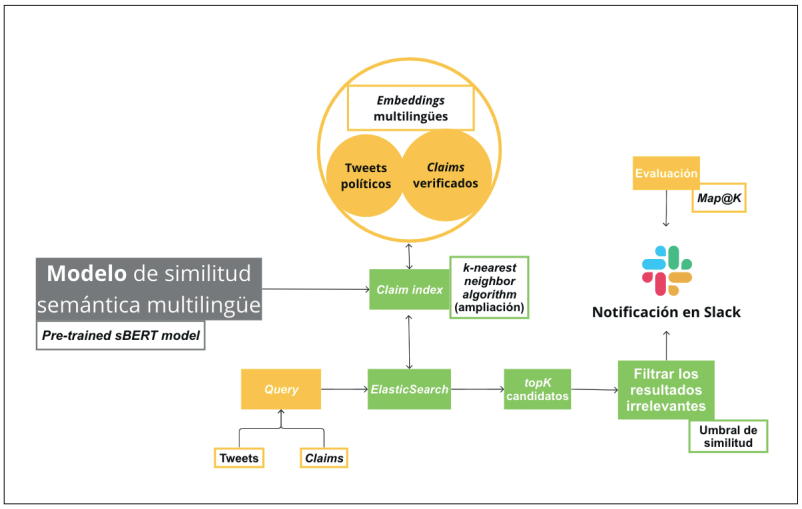

Figura 2. Arquitectura del modelo con un sistema de búsqueda semántica y un umbral

Al igual que en el experimento anterior, el primer paso fue entrenar un modelo de búsqueda semántica pre-entrenado, y en un segundo paso, entrenar un clasificador binario de paráfrasis utilizando el conjunto de datos de entrenamiento etiquetados para este propósito. Ambos modelos se integraron en un prototipo que permite primero la búsqueda semántica y, en una segunda fase, identificar la paráfrasis con el sistema de clasificación.

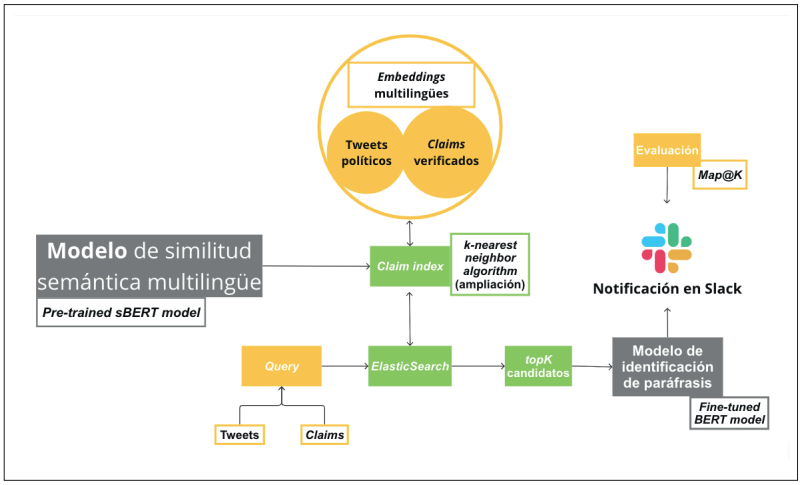

Figura 3. Arquitectura del modelo con un sistema de búsqueda semántica y clasificador

Como se muestra en la figura 3, el prototipo se basa en un generador de *embeddings* a partir de afirmaciones de políticos en *Twitter* y de la hemeroteca construida con las verificaciones publicadas previamente (*Claim index*). A partir de los nuevos tweets y afirmaciones (*Query*) se realiza una búsqueda (*ElasticSearch*) en los *embeddings* y se recuperan los resultados más acertados (*topK candidates*). Para filtrar los *topK candidates*, se utiliza el modelo de paráfrasis, que selecciona los resultados que arrojan un mejor emparejamiento (*Paraphrasis identification model*) y se envía una alerta al programa de mensajería *Slack*.

La retroalimentación de los periodistas en *Slack*, que etiquetan si los candidatos escogidos por la herramienta son similares o no, permite hacer una evaluación del modelo en el mundo real utilizando la precisión media a diferentes valores de K (MAP@K).

Para comprobar cuál era la mejor opción, realizamos una evaluación MAP@K en cada una de las tres arquitecturas. Consiste en evaluar cuantitativamente el porcentaje de candidatos *topK* recuperados, es decir, la precisión del sistema. A nivel individual, en el primer y tercer caso se realiza la evaluación del clasificador, que se centró en la precisión (porcentaje de registros correctamente clasificados), el retorno (porcentaje de registros similares que son devueltos) y la puntuación F1 (una métrica que resume las dos previas).

5.3. Resultados de los experimentos

La tercera estrategia (búsqueda semántica + clasificador) resultó ser la más exitosa para nuestro caso de uso. En cuanto al modelo de clasificación de similitud, realizamos algunas pruebas para elegir una de las versiones pre-entrenadas de modelos tipo *BERT* que ofrece *Huggingface*, que cuenta con una de las bibliotecas más populares y utilizadas en el campo del lenguaje natural. Entre estas pruebas, los modelos con mejor rendimiento corresponden a *microsoft/Multilingual-MiniLM-L12-v2* y a *xlm-roberta-base*.

Tabla 3. Resultados preliminares del ensayo con distintos modelos pre-entrenados

Modelos - 2checks entrenamiento	Conjunto de entrenamiento	Umbral	Conjunto	Similitud %	Precisión	*Recall*	*F1-score*	*Accuracy*
microsoft/Multilingual-MiniLM-L12-H384	2checks (66% similar)	0,5	Test MRPC	66%	67,32%	98,86%	80,10%	67,35%
			Test Benchmark - ClaimHunter	45%	79,45%	71,32%	75,16%	78,05%
			Test PAWSX-es	44%	44,79%	99,89%	61,84%	44,82%
xlm-roberta-base (bce - preprocess tweets)	2checks (66% similar)	0,5	Test MRPC	66%	67,36%	99,21%	80,24%	67,52%
			Test Benchmark - ClaimHunter	45%	80,88%	69,98%	75,04%	78,32%
			Test Benchmark - ClaimHunter - only Sentences (Set6)	60%	92,14%	50,96%	65,62%	67,58%
			Test PAWSX-es	44%	44,76%	99,89%	61,82%	44,77%
xlm-roberta-base (bce - preproc-assign-labels -current ml commons version)	2checks (66% similar)	0,5	Test MRPC	66%	67,46%	98,77%	80,17%	96,18%
			Test Benchmark - ClaimHunter	45%	79,59%	67,11%	72,82%	76,67%
			Test PAWSX-es	44%	44,79%	100,00%	61,87%	44,82%

Modelos - 2checks entrenamiento	Conjunto de entrenamiento	Umbral	Conjunto	Similitud %	Precisión	*Recall*	*F1-score*	*Accuracy*
xlm-roberta-base (bce - preproc-assign-labels -local)	2checks (66% similar)	0,5	Test MRPC	66%	66,90%	98,77%	79,77%	66,71%
			Test Benchmark - ClaimHunter	45%	78,38%	66,54%	71,98%	75,87%
			Test PAWSX-es	44%	44,79%	100,00%	61,87%	44,82%
xlm-roberta-base (bce - preprocess tweets)	2checks (66% similar)	ROC optimal = 0,331424	Test MRPC	66%	67,35%	98,25%	79,91%	67,17%
			Test Benchmark - ClaimHunter	45%	75,52%	75,81%	75,67%	77,29%
			Test PAWSX-es	44%	44,82%	100,00%	61,89%	44,87%
microsoft/mdeberta-v3-base (bce - preprocess tweets)	2checks (66% similar)	0,5	Test MRPC	66%	66,98%	99,82%	80,17%	67,17%
			Test Benchmark - ClaimHunter	45%	66,44%	92,73%	77,41%	74,80%
			Test PAWSX-es	44%	44,77%	100,00%	61,85%	44,77%
microsoft/mdeberta-v3-base (bce - preprocess-tweets) adjustment of deberta specific parameters (warmup steps, epsilon, weight decay)	2checks (66% similar)	0,5	Test MRPC	66%	67,24%	99,74%	99,69%	67,52%
			Test Benchmark - ClaimHunter	45%	66,17%	92,73%	77,23%	74,53%
			Test PAWSX-es	44%	44,77%	100,00%	61,85%	44,77%
microsoft/mdeberta-v3-base (bce - preprocess-tweets) adjustment of deberta specific parameters (warmup steps, epsilon, weight decay)	2checks (66% similar)	ROC optimal = 0,998581	Test MRPC	66%	67,80%	99,21%	80,55%	68,18%
			Test Benchmark - ClaimHunter	45%	78,55%	80,88%	79,70%	80,81%
			Test PAWSX-es	44%	44,79%	100,00%	61,87%	44,82%
bert-base-multilingual-cased (bce - preprocess-tweets)	2checks (66% similar)	0,5	Test MRPC	66%	66,71%	99,82%	79,97%	66,76%
			Test Benchmark - ClaimHunter	45%	71,76%	76,29%	73,96%	74,98%
			Test PAWSX-es	44%	44,79%	100,00%	61,87%	44,82%
sentence-transformers/ paraphrase-multilingual-mpnet-base-v2 (bce - preprocess-tweets)	2checks (66% similar)	0,5	Test MRPC	66%	66,39%	99,30%	79,58%	66,12%
			Test Benchmark - ClaimHunter	45%	56,22%	82,89%	67,03%	61,98%
			Test PAWSX-es	44%	44,79%	100,00%	61,87%	44,82%
sentence-transformers/ stsb-xlm-r-multilingual (bce - preprocess-tweets)	2checks (66% similar)	0,5	Test MRPC	66%	67,53%	93,43%	78,40%	65,77%
			Test Benchmark - ClaimHunter	45%	66,97%	69,79%	68,35%	69,90%
			Test PAWSX-es	44%	44,79%	99,89%	61,84%	44,82%
sentence-transformers/ paraphrase-xlm-r-multilingual-v1 (bce - preprocess-tweets)	2checks (66% similar)	0,5	Tesr MRPC	66%	66,82%	90,46%	76,86%	63,80%
			Test Benchmark - ClaimHunter	45%	62,49%	69,12%	65,64%	66,30%
			Test PAWSX-es	44%	44,81%	99,89%	61,87%	44,87%

Ambos modelos ofrecieron un buen rendimiento en precisión y recuperación, pero elegimos el *microsoft/Multilingual-MiniLM-L12-v2* por ser un modelo más ligero que lo hacía más adecuado para su uso en la redacción.

En los experimentos se utilizó el modelo *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2* para comprobar la recuperación de candidatos. Se evaluó la eficacia del modelo al agregar un filtro adicional utilizando el modelo *microsoft/Multilingual-MiniLM-L12-H384*. Para medir el rendimiento, se utilizaron tres métricas muy comunes en los sistemas de recuperación de información: *MAP@K*, *Recall@K* y *Mean Reciprocal Ranking* (*MRR*). El objetivo fue evaluar la capacidad de cada modelo para recuperar los resultados más relevantes de acuerdo con los criterios de búsqueda del usuario.

Al comparar los resultados del experimento, se pudo confirmar que el filtrado del modelo con la tercera arquitectura contribuye a recuperar candidatos más adecuados en las primeras posiciones (tabla 5). Esto se traduce en una mayor precisión de los primeros candidatos recuperados, tal como se puede observar en los valores superiores del indicador *MAP@K* para valores más pequeños de K.

Tabla 4. Resultados de los ensayos del modelo de clasificación

Similitud semántica - Modelo de clasificación													
Modelo	Dataset de entrenamiento	Distribución entrenamiento	Tamaño	Dataset de test	Test size	Distribución test	Precisión		Recall		F1		Accuracy
							class 0	class 1	class 0	class 1	class 0	class 1	
xlm-roberta-base	*Fact Check Explorer* y otros	66% similar	7.762	Test Benchmark	2.246	45% similar	77,2%	80,0%	84,5%	71,3%	80,7%	75,4%	78,4%
				Prodigy Benchmark	7.443	31% similar	85,4%	67,0%	84,4%	68,8%	84,9%	67,9%	79,5%
microsoft/ Multilingual-MiniLM-L12-H384	*Fact Check Explorer* y otros	66% similar	7.762	Test Benchmark	2.246	45% similar	77%	79,4%	83,9%	71,3%	80,3%	75,2%	78,0%
				Prodigy Benchmark	7.443	31% similar	86,3%	64,7%	82,0%	71,8%	84,1%	68,1%	78,8%

Tabla 5. Promedio de resultados de los tres experimentos

Evaluación de la recuperación de candidatos (media de las tres arquitecturas)				
	K = 1	K = 3	K = 5	K = 10
MAP@K	0,7471	0,6971	0,6842	0,6864
Recall@K	0,3383	0,5615	0,6646	0,7836
MRR	0,8528			

Evaluación de la recuperación de candidatos (+ modelo de filtrado) (media de las tres arquitecturas)				
	K = 1	K = 3	K = 5	K = 10
MAP@K	0,8237	0,7122	0,6837	0,6646
Recall@K	0,3638	0,5625	0,6392	0,7148
MRR	0,9169			

6. Discusión y futuras líneas de investigación

El primer hallazgo de la investigación indica que encontrar un equilibrio entre la precisión del sistema y su capacidad para recuperar todas las frases similares resulta una tarea desafiante. A medida que se aumenta la tasa de recuperación, la precisión en las recomendaciones proporcionadas a los verificadores disminuye, lo que repercute negativamente en su nivel de confianza en el algoritmo y en el esfuerzo necesario para revisar la información. Dado que, en general, el volumen de frases no relacionadas es significativamente mayor que el de frases similares, el objetivo es optimizar la precisión del sistema con el propósito de evitar una alta incidencia de falsos positivos.

Se contempla otro problema en el proceso, relacionado con la disminución de la velocidad de búsqueda a medida que aumenta la base de verificaciones. Esto ocurre si los modelos utilizados no escalan linealmente. Por esta razón, se ha optado por seleccionar modelos más rápidos en detrimento de modelos más pesados, siempre y cuando la mejora en el rendimiento que proporcionen estos últimos sea marginal, es decir, no superior al 3%. En situaciones reales, donde el número de pares de frases a comparar puede alcanzar cifras de millones, es fundamental priorizar la escalabilidad por encima de la precisión.

En los experimentos se comprueba también que se enfrenta un desafío significativo en la falta de contexto temporal y espacial. En el futuro, se hace necesario, por lo tanto, diseñar una estrategia que permita generar y almacenar metadatos relevantes asociados a cada tweet y frase en el sistema. La recuperación de candidatos deberá considerar tanto la relevancia semántica de las frases como la concordancia de sus metadatos espaciales y temporales. Para resolver el problema temporal en el sistema actual, se sugiere la utilización de ventanas temporales para recuperar los *topK* candidatos válidos de forma efectiva. Aunque es una solución sencilla, esta estrategia resulta ser eficaz en la recuperación de información adecuada en el marco de la arquitectura propuesta.

La identificación precisa de entidades (*entity linking*) representa otro desafío difícil de abordar para los sistemas de *claim matching* en el contexto político. Dado que el modelo solo cuenta con la información de la propia frase, es incapaz de interpretar que términos y expresiones como “Feijóo”, “el presidente de los populares” o “el líder de la oposición” hacen referencia a la misma persona, tal como se expone en el apartado 4.2. Si bien en algunos casos específicos, el uso de aproximaciones basadas en diccionarios puede reducir este problema, se requiere la creación de soluciones más generales.

7. Conclusiones

Los resultados del estudio indican que los sistemas de similitud semántica son una herramienta valiosa para detectar paráfrasis en el discurso político y mejorar la eficacia de los modelos de *claim matching* que contribuyen al fact-checking automatizado. En concreto, la com-

> “Los sistemas de similitud semántica son una herramienta valiosa para detectar paráfrasis en el discurso político y mejorar la eficacia del *claim matching* en el fact-checking automatizado”

binación de sistemas de búsquedas semánticas y clasificadores para la recuperación de verificaciones previas puede mejorar la eficiencia y efectividad del *claim matching*, permitiendo a las organizaciones de fact-checking detectar afirmaciones ya verificadas más rápidamente y con mayor precisión.

> La combinación de sistemas de búsquedas semánticas y clasificadores puede mejorar la eficiencia y efectividad del *claim matching*, permitiendo detectar afirmaciones ya verificadas de manera más rápida y precisa

La conclusión general que se puede extraer de los resultados del experimento es que el filtrado del modelo utilizando este tipo de arquitectura puede ser altamente beneficioso para mejorar la precisión de los primeros candidatos recuperados. La evidencia proporcionada por los experimentos muestra que, al utilizar esta técnica, se pueden recuperar candidatos de frases similares más adecuados y situarlos en las primeras posiciones, ganando en agilidad sin perder precisión, por lo que puede ser una opción especialmente efectiva para el fact-checking. Aunque este resultado puede tener importantes implicaciones prácticas en diversas áreas.

El programa *ClaimCheck* emplea un enfoque basado en el modelo *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2*, que combina la búsqueda semántica con un clasificador para filtrar los resultados obtenidos por medio del modelo *microsoft/Multilingual-MiniLM-L12-H384*. Los resultados obtenidos por *ClaimCheck* son superiores a los de otros modelos evaluados, como se evidencia en las métricas de *MAP@K*, *Recall@K* y *Mean Reciprocal Ranking*.

La experiencia de *ClaimCheck* muestra las vías de experimentación más exitosas con resultados positivos para la redacción de *Newtral*, aunque también expone algunos problemas que requieren de solución para mejorar la eficacia del modelo. Entre ellos se destacan la falta de contexto, el reconocimiento preciso de entidades y la necesidad de equilibrar la precisión con la recuperación de todas las frases similares.

El diseño experimental presentado en este artículo puede ser utilizado como punto de partida para futuras investigaciones en el campo del fact-checking automatizado y el uso de sistemas de similitud semántica en la identificación de afirmaciones falsas en el discurso político, no solo para extraer verificaciones previas y poder reutilizarlas de forma más rápida sin tener que duplicar el trabajo, sino también para el fact-checking en directo o para crear sistemas de alerta ante campañas de desinformación. Asimismo, el presente estudio sugiere posibles direcciones para investigaciones futuras que permitan solucionar los desafíos presentes en el modelo propuesto, y mejorar su capacidad de ser adoptado por organizaciones especializadas en fact-checking.

8. Referencias

Adair, Bill (2021). "The lessons of Squash, Duke's automated fact-checking platform". *Poynter*, 16 June.
https://www.poynter.org/fact-checking/2021/the-lessons-of-squash-the-first-automated-fact-checking-platform

Adair, Bill; **Li, Chengkai**; **Yang, Jun**; **Yu, Cong** (2018). *Automated pop-up fact-checking: challenges & progress*.
https://ranger.uta.edu/~cli/pubs/2019/popupfactcheck-cj19-adair.pdf

Agadjanian, Alexander; **Bakhru, Nikita**; **Chi, Victoria**; **Greenberg, Devyn**; **Hollander, Byrne**; **Hurt, Alexander**; **Kind, Joseph**; **Lu, Ray**; **Ma, Annie**; **Nyhan, Brendan**; **Pham, Daniel**; **Qian, Michael**; **Tan, Mackinley**; **Wang, Clara**; **Wasdahl, Alexander**; **Woodruff, Alexandra** (2019). "Counting the Pinocchios: the effect of summary fact-checking data on perceived accuracy and favorability of politicians". *Research & politics*, v. 6, n. 3.
https://doi.org/10.1177/2053168019870351

Arslan, Fatma (2021). *Modeling factual claims with semantic frames: definitions, datasets, tools, and fact-checking applications*. Doctoral dissertation. The University of Texas at Arlington.
https://rc.library.uta.edu/uta-ir/bitstream/handle/10106/30765/ARSLAN-DISSERTATION-2021.pdf

Babakar, Mevan; **Moy, Will** (2016). *The state of automated factchecking. How to make factchecking dramatically more effective with technology we have now*. Full Fact.
https://fullfact.org/media/uploads/full_fact-the_state_of_automated_factchecking_aug_2016.pdf

Baker, Collin F.; **Fillmore, Charles J.**; **Lowe, John B.** (1998). "The Berkeley FrameNet project". In: *Proceedings of the joint conference of the international conference on computational linguistics and the Association for Computational Linguistics (Coling-ACL)*, pp. 86-90.
https://aclanthology.org/C98-1013.pdf

Beltrán, Javier; **Míguez, Rubén**; **Larraz, Irene** (2019). "ClaimHunter: an unattended tool for automated claim detection on *Twitter*". *KnOD@WWW. CEUR workshop proceedings*, v. 2877, n. 3.
https://ceur-ws.org/Vol-2877/paper3.pdf

Corney, David (2021a). "How does automated fact checking work?". *Full Fact*, 5 July.
https://fullfact.org/blog/2021/jul/how-does-automated-fact-checking-work

Corney, David (2021b). "Towards a common definition of claim matching". *Full Fact*, 5 October.
https://fullfact.org/blog/2021/oct/towards-common-definition-claim-matching

Dolan, William B.; **Brockett, Chris** (2005). "Automatically constructing a corpus of sentential paraphrases". In: *Proceedings of the third international workshop on paraphrasing* (*IWP2005*), pp. 9-16.
https://aclanthology.org/I05-5002.pdf

Floodpage, Sebastien (2021). "How fact checkers and *Google.org* are fighting misinformation". *Google*, 31 March.
https://blog.google/outreach-initiatives/google-org/fullfact-and-google-fight-misinformation

Graves, Lucas (2018). *Understanding the promise and limits of automated fact-checking*. Reuters Institute for the Study of Journalism. Factsheets.
https://ora.ox.ac.uk/objects/uuid:f321ff43-05f0-4430-b978-f5f517b73b9b

Hassan, Aumyo; **Barber, Sarah J.** (2021). "The effects of repetition frequency on the illusory truth effect". *Cognitive research: principles and implications*, v. 6, n. 38.
https://doi.org/10.1186/s41235-021-00301-5

Hassan, Naeemul; **Adair, Bill**; **Hamilton, James T.**; **Li, Chengkai**; **Tremayne, Mark**; **Yang, Jun**; **Yu, Cong** (2015). "The quest to automate fact-checking". In: *Proceedings of the 2015 computation + journalism symposium*. Columbia University.
http://cj2015.brown.columbia.edu/papers/automate-fact-checking.pdf

Hassan, Naeemul; **Arslan, Fatma**; **Li, Chengkai**; **Tremayne, Mark** (2017). "Toward automated fact-checking: detecting check-worthy factual claims by ClaimBuster". In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (*KDD '17*). New York: Association for Computing Machinery, pp. 1803-1812.
https://doi.org/10.1145/3097983.3098131

Hövelmeyer, Alica; **Boland, Katarina**; **Dietze, Stefan** (2022). "SimBa at CheckThat! 2022: lexical and semantic similarity based detection of verified claims in an unsupervised and supervised way". In: *CLEF 2022: Conference and labs of the evaluation forum*, 5-8 September, Bolonia, Italia.
https://ceur-ws.org/Vol-3180/paper-40.pdf

Jiang, Ye; **Song, Xingyi**; **Scarton, Carolina**; **Aker, Ahmet**; **Bontcheva, Kalina** (2021). "Categorising fine-to-coarse grained misinformation: an empirical study of Covid-19 Infodemic". *Arxiv*.
https://doi.org/10.48550/arXiv.2106.11702

Kazemi, Ashkan; **Garimella, Kiran**; **Gaffney, Devin**; **Hale, Scott A.** (2021). "Claim matching beyond English to scale global fact-checking". In: *Proceedings of the 59th Annual meeting of the Association for Computational Linguistics and the 11th International joint conference on natural language processing*. Association for Computational Linguistics, pp. 4504-4517.
https://doi.org/10.18653/v1/2021.acl-long.347

Kazemi, Ashkan; **Li, Zehua**; **Pérez-Rosas, Verónica**; **Hale, Scott A.**; **Mihalcea, Rada** (2022). "Matching tweets with applicable fact-checks across languages". *Arxiv*.
https://doi.org/10.48550/arXiv.2202.07094

Kessler, Glenn; **Fox, Joe** (2021). "The false claims that Trump keeps repeating". *The Washington Post*, 20 January.
https://www.washingtonpost.com/graphics/politics/fact-checker-most-repeated-disinformation

Lan, Zhenzhong; **Chen, Mingda**; **Goodman, Sebastian**; **Gimpel, Kevin**; **Sharma, Piyush**; **Soricut, Radu** (2020). "ALBERT: a lite Bert for self-supervised learning of language representations". In: *Conference paper at International conference on learning representations (ICLR)*. *Arxiv*.
https://doi.org/10.48550/arXiv.1909.11942

Lim, Chloe (2018). "Checking how fact-checkers check". *Research & politics*, v. 5, n. 3.
https://doi.org/10.1177/2053168018786848

Mansour, Watheq; **Elsayed, Tamer**; **Al-Ali, Abdulaziz** (2022). "Did I see it before? Detecting previously-checked claims over *Twitter*". *Lecture notes in computer science*, pp. 367-381.
https://doi.org/10.1007/978-3-030-99736-6_25

Martín, Alejandro; **Huertas-Tato, Javier**; **Huertas-García, Álvaro**; **Villar-Rodríguez, Guillermo**; **Camacho, David** (2021). "FacTeR-check: semi-automated fact-checking through semantic similarity and natural language inference". *Arxiv*.
https://doi.org/10.48550/arXiv.2110.14532

Mukherjee, Amit; **Sela, Eitan**; **Al-Saadoon, Laith** (2020). "Building an NLU-powered search application with *Amazon SageMaker* and the *Amazon opensearch* service KNN feature". *Amazon SageMaker, artificial intelligence*, 26 October.
https://aws.amazon.com/es/blogs/machine-learning/building-an-nlu-powered-search-application-with-amazon-sagemaker-and-the-amazon-es-knn-feature

Murray, Samuel; **Stanley, Matthew**; **McPhetres, Jon**; **Pennycook, Gordon**; **Seli, Paul** (2020). "'I've said it before and I will say it again…': repeating statements made by Donald Trump increases perceived truthfulness for individuals across the political spectrum". *PsyArXiv preprints*, 15 January.
https://doi.org/10.31234/osf.io/9evzc

Nakov, Preslav; **Corney, David**; **Hasanain, Maram**; **Alam, Firoj**; **Elsayed, Tamer**; **Barrón-Cedeño, Alberto**; **Papotti, Paolo**; **Shaar, Shaden**; **Da-San-Martino, Giovanni** (2021). "Automated fact-checking for assisting human fact-checkers". *International joint conference on artificial intelligence*. *Arxiv*.
https://doi.org/10.48550/arXiv.2103.07769

Nakov, Preslav; **Da-San-Martino, Giovanni**; **Alam, Firoj**; **Shaar, Shaden**; **Mubarak, Hamdy**; **Babulkov, Nikolay** (2022). "Overview of the CLEF-2022 CheckThat! Lab task 2 on detecting previously fact-checked claims". In: *CLEF 2022: conference and labs of the evaluation forum*, 5-8 septiembre, Bolonia, Italia.
https://ceur-ws.org/Vol-3180/paper-29.pdf

Nguyen, Vincent; **Karimi, Sarvnaz**; **Xing, Zhenchang** (2021). "Combining shallow and deep representations for text-pair classification". In: *Proceedings of the 19th Annual workshop of the Australasian Language Technology Association*, pp. 68-78.
https://aclanthology.org/2021.alta-1.7.pdf

Phillips, Whitney (2018). *The oxygen of amplification. Better pratices for reporting on extremists, antagonists, and manipulators online*. Data & Society Research Institute.
https://datasociety.net/wp-content/uploads/2018/05/FULLREPORT_Oxygen_of_Amplification_DS.pdf

Porter, Ethan; **Wood, Thomas J.** (2021). "The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom". *Proceedings of the National Academy of Sciences of the United States of America*, v. 118, n. 37.
https://doi.org/10.1073/pnas.2104235118

Real, Andrea (2021). "Casado mezcla diferentes estadísticas de empleo para asegurar que hay 4 millones de parados, pero es falso". *Newtral*, 6 octubre.
https://www.newtral.es/parados-espana-casado-pp-factcheck/20211007

Reimers, Nils; **Gurevych, Iryna** (2019). "Sentence-bert: sentence embeddings using siamese bert-networks". In: *Proceedings of the 2019 Conference on empirical methods in natural language processing and the 9th International joint conference on natural language processing* (*EMNLP-IJCNLP*). Hong Kong, November, pp. 3982-3992.
https://doi.org/10.18653/v1/D19-1410

Shaar, Shaden; **Alam, Firoj**; **Da-San-Martino, Giovanni**; **Nakov, Preslav** (2021a). "The role of context in detecting previously fact-checked claims". *Arxiv*.
https://doi.org/10.48550/arXiv.2104.07423

Shaar, Shaden; **Babulkov, Nikolay**; **Da-San-Martino, Giovanni**; **Nakov, Preslav** (2020). "That is a known lie: detecting previously fact-checked claims". In: *Proceedings of the 58th Annual meeting of the Association for Computational Linguistics*, pp. 3607-3618.
https://doi.org/10.18653/v1/2020.acl-main.332

Shaar, Shaden; **Haouari, Fatima**; **Mansour, Watheq**; **Hasanain, Maram**; **Babulkov, Nikolay**; **Alam, Firoj**; **Da-San-Martino, Giovanni**; **Elsayed, Tamer**; **Nakov, Preslav** (2021b). "Overview of the CLEF-2021 CheckThat! Lab task 2 on detecting previously fact-checked claims in tweets and political debates". In: *CLEF 2021: Conference and labs of the evaluation forum*, 21-24 September, Bucharest, Romania.
https://ceur-ws.org/Vol-2936/paper-29.pdf

Sheng, Qiang; **Cao, Juan**; **Zhang, Xueyao**; **Li, Xirong**; **Zhong, Lei** (2021). "Article reranking by memory-enhanced key sentence matching for detecting previously fact-checked claims". In: *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).
https://doi.org/10.18653/v1/2021.acl-long.425

Sippitt, Amy (2020). *What is the impact of fact checkers' work on public figures, institutions and the media?*. Africa Check, Chequeado and Full Fact.
https://fullfact.org/media/uploads/impact-fact-checkers-public-figures-media.pdf

Stanford Institute for Human-Centered Artificial Intelligence (2023). *Artificial intelligence index*. Stanford University.
https://aiindex.stanford.edu/report

The Washington Post (2018). "Meet the bottomless Pinocchio | Fact Checker". [Video]. *YouTube*, 10 December.
https://www.youtube.com/watch?v=zoS1sVZRfUU

Thorne, James; **Vlachos, Andreas** (2018). "Automated fact checking: task formulations, methods and future directions". *Arxiv*.
https://doi.org/10.48550/arXiv.1806.07687

Wardle, Claire (2018). "Lessons for reporting in an age of disinformation". *Medium*, 28 December.
https://medium.com/1st-draft/5-lessons-for-reporting-in-an-age-of-disinformation-9d98f0441722

Zeng, Xia; **Abumansour, Amani S.**; **Zubiaga, Arkaitz** (2021). "Automated fact-checking: a survey". *Language and linguistics compass*, v. 15, n. 10.
https://doi.org/10.1111/lnc3.12438