Protocolo metodológico para el desarrollo de análisis de contenido asistido por inteligencia artificial fiable y válido: guía práctica con *ChatGPT*

Methodological protocol for developing reliable and valid AI-assisted content analysis: a practical guide with *ChatGPT*

Manuel Goyanes; **Luis De-Marcos**

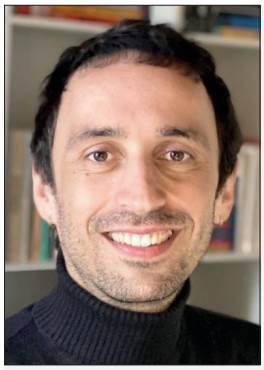
Goyanes, Manuel; **De-Marcos, Luis** (2025). "Protocolo metodológico para el desarrollo de análisis de contenido asistido por inteligencia artificial fiable y válido: guía práctica con ChatGPT". *Anuario ThinkEPI*, v. 19, e19a03.

https://doi.org/10.3145/thinkepi.2025.e19a07

Publicado en *IweTel* el 25 de mrazo de 2025

Manuel Goyanes
https://www.directorioexit.info/ficha3719
https://orcid.org/0000-0001-8329-0610
Universidad Carlos III de Madrid
Departamento de Comunicación
manuel.goyanes@uc3m.es

Luis De-Marcos
https://orcid.org/0000-0003-0718-8774
Universidad de Alcalá
Departamento de Ciencias de la Computación
luis.demarcos@uah.es



Resumen: Desde el surgimiento de la inteligencia artificial generativa, siendo *ChatGPT* uno de sus principales exponentes, la codificación automatizada de texto mediante análisis de contenido se ha visto considerablemente facilitada. Sin embargo, a pesar de estos avances, aún no se dispone de un protocolo metodológico que analice esta adaptación y ofrezca recomendaciones para un uso riguroso de la inteligencia artificial en el análisis de contenido. El presente estudio propone una guía metodológica de fácil aplicación para la implementación de análisis de contenido automatizado con inteligencia artificial, con el objetivo de garantizar la rigurosidad en términos de fiabilidad y validez de los datos obtenidos. Asimismo, se busca que dicho protocolo sea práctico y accesible para investigadores sin formación especializada en programación o ciencias sociales computacionales, razonando sobre los procesos más relevantes.

Palabras Clave: Análisis de contenido; Inteligencia artificial; *ChatGPT*; Protocolo metodológico.

Abstract: Since the emergence of generative artificial intelligence, with *ChatGPT* as one of its leading examples, the automated coding of text through content analysis has become significantly more accessible. However, despite these advancements, there is still no established methodological protocol that examines this adaptation and provides recommendations for the rigorous use of artificial intelligence in content analysis. This study proposes a user-friendly methodological guide for implementing AI-assisted automated content analysis, aiming to ensure the reliability and validity of the data obtained. Furthermore, the protocol is designed to be practical and accessible to researchers without specialized training in programming or computational social sciences, offering reasoned guidance on the most relevant processes.

Keywords: Content analysis; Artificial intelligence; *ChatGPT*; Methodological protocol.

El análisis de contenido ha sido históricamente una de las técnicas de investigación más relevantes en el campo de la comunicación (**Riffe** et al., 2019). En términos generales, el análisis de contenido se centra en el estudio de la comunicación humana registrada, que generalmente implica la codificación del contenido para identificar la presencia de ciertos rasgos, categorías o significados. Este tipo de análisis permite relacionar la aparición de elementos codificados con otros factores, como las características del emisor o los efectos en el receptor. Además, se aplica al estudio de diversas fuentes, como revistas, periódicos, transcripciones, discursos, o revistas científicas, entre otros (**Hayes**; **Krippendorff**, 2007). Tradicionalmente, el análisis de contenido se ha venido efectuando de forma manual, en donde codificadores (jueces) independientes efectúan inferencias del contenido para ofrecer resultados descriptivos o relacionales de las unidades de análisis examinadas (**Goyanes**; **Piñeiro-Naval**, 2024).

En los últimos años, diversas herramientas computacionales han automatizado en gran medida la codificación de grandes volúmenes de contenido, reduciendo así la carga de trabajo de los investigadores humanos y los costos asociados a la codificación manual (**Goyanes** *et al.*, 2024). Más allá de este desarrollo, la inteligencia artificial, con modelos como *ChatGPT*, ha supuesto un cambio disruptivo en este tipo de análisis, impulsando el desarrollo de estudios más ambiciosos y optimizando el procesamiento y análisis de datos masivos. Este fenómeno ha dado lugar a un creciente número de investigaciones que examinan el impacto de *ChatGPT* en el análisis de contenido, examinando su fiabilidad para realizar inferencias en diferentes modalidades de análisis, (**Goyanes** *et al.*, 2024; **Maj** *et al.*, 2025). Sin embargo, a pesar del continuo avance en la automatización y en la evaluación de la fiabilidad de la inteligencia artificial, la comunidad científica aún carece de un protocolo estandarizado que estructure este proceso asistido por inteligencia artificial. En este contexto, el presente estudio se articula en torno a la siguiente pregunta de investigación: ¿cómo llevar a cabo un análisis de contenido riguroso asistido por *ChatGPT*? Para responder a esta cuestión, se lleva a cabo un protocolo metodológico que pueda ser implementado por investigadores sin conocimientos avanzados en computación.

Paso 1. Documentación del proceso de análisis con *ChatGPT*

El primer paso del protocolo consiste en documentar cada fase del procedimiento de análisis con *ChatGPT*, lo que implica seguir tres procedimientos:

Registrar los *prompt*s utilizados para el análisis de contenido

El registro detallado de las instrucciones o *prompts* empleados en el análisis de contenido es fundamental para garantizar la trazabilidad del proceso y la reproducibilidad del estudio. Esto implica documentar de manera explícita las instrucciones proporcionadas a la inteligencia artificial, ya sea en el apartado metodológico del trabajo o en un apéndice del artículo. Dado que la respuesta generativa de los modelos de inteligencia artificial está significativamente condicionada por la formulación del *prompt*, su documentación permite analizar cómo la IA interpreta y procesa los datos. Además de contribuir a la transparencia del estudio, este registro facilita la identificación de posibles sesgos inherentes a la formulación de las instrucciones, lo que facilita la evaluación del análisis realizado.

Especificar la versión de *ChatGPT* empleada

Para garantizar una mayor transparencia en el estudio, es fundamental especificar el modelo de *ChatGPT* empleado, ya que cada versión presenta variaciones en su arquitectura que pueden influir en las inferencias generadas.

Especificar el preprocesamiento de datos efectuado

Antes de procesar el contenido en la plataforma de *ChatGPT*, muchos investigadores aplican técnicas de preprocesamiento para optimizar la calidad de los datos. Este proceso puede incluir la eliminación de redundancias, caracteres especiales o elementos con escaso valor informativo para la generación de inferencias. En términos metodológicos, la limpieza y estructuración de la base de datos facilita el análisis realizado por la inteligencia artificial y, por lo tanto, puede influir en los resultados obtenidos. Es por ello por lo que debe explicitarse el preprocesamiento de datos efectuado.

Paso 2. Explicación del libro de códigos y su aplicación en la inferencia

El siguiente paso en el protocolo metodológico para el uso de inteligencia artificial consiste en la definición del libro de códigos y la alineación entre el entrenamiento humano y las instrucciones proporcionadas a *ChatGPT*.

Definición del libro de códigos

En el libro de códigos se definen las variables objeto de estudio y su correspondiente codificación a través de categorías preestablecidas. Por ejemplo, en un libro de códigos diseñado para un análisis de sentimiento, es fundamental incluir una definición clara del concepto de análisis de sentimiento, así como la descripción detallada de las tres categorías de clasificación: positivo, negativo y neutro.

Alineación entre entrenamiento humano y los *prompts* de *ChatGPT* para realizar inferencias

A continuación, es necesario establecer reglas claras sobre cómo asignar cada categoría a las observaciones, es decir, ofrecer una descripción detallada de los criterios de codificación. Dado que en este proceso participan tanto la inteligencia artificial como un codificador humano, es fundamental incluir explícitamente los criterios para la realización de inferencias. En este sentido, los *prompts* proporcionados a *ChatGPT* deben reflejar con precisión los criterios utilizados por los codificadores humanos y viceversa, garantizando así la coherencia en la asignación de categorías. Para lograrlo, es recomendable utilizar instrucciones que permitan a la IA aplicar las categorías de manera alineada con el criterio humano. Con el objetivo de aumentar la transparencia del estudio, se recomienda explicitar los criterios de codificación para el entrenamiento humano y los *prompts* para *ChatGPT*.

Paso 3. Análisis de consistencia y secuenciación de la codificación

El tercer paso consiste en la realización de un análisis de la consistencia interna de las inferencias de *ChatGPT* y una descripción detallada de la secuencia de codificación efectuada. Esto se traduce en los siguientes procedimientos:

Evaluar la consistencia interna de la inteligencia artificial

Este paso puede no ser esencial cuando el análisis de *prompts* se limita a inferencias simples, es decir, cuando se centra en el contenido manifiesto en lugar del contenido latente. Aunque la literatura previa (**Jiang**; **Gao**; **Karniadakis**, 2025; **Moon**, 2025; **Mervaala**; **Kousa**, 2025) sugiere que *ChatGPT* presenta mayores dificultades en tareas de análisis de contenido manifiesto (por ejemplo, en el conteo preciso de palabras dentro de un texto), la evaluación de la consistencia interna resulta fundamental para responder a la siguiente pregunta: ¿responde *ChatGPT* de manera uniforme ante los mismos *prompts*? Este procedimiento permite evaluar de manera sencilla la fiabilidad del modelo, determinando si genera respuestas significativamente distintas ante instrucciones idénticas. Para llevar a cabo esta evaluación, se pueden formular los mismos *prompts* en múltiples ocasiones y comparar las respuestas obtenidas, lo que facilita la identificación de posibles variaciones en la inferencia generada por la IA. En caso de divergencia significativa en los resultados obtenidos para

distintas formulaciones con idéntico *prompt*, debe aplicarse algún procedimiento estadístico, como el intervalo de confianza, realizando múltiples ejecuciones que permitan analizar la variabilidad y estabilidad del resultado.

Secuenciación del proceso de codificación

Este procedimiento consiste en definir la secuencia en la que se lleva a cabo la codificación del contenido entre los distintos codificadores involucrados en el análisis de contenido. En la actualidad, las herramientas más utilizadas para la codificación, tanto en enfoques manuales como automatizados, incluyen modelos de inteligencia artificial generativa como *ChatGPT*, codificadores humanos y otras herramientas computacionales especializadas, como *BERT*. La elección del orden o secuencia en la codificación es un aspecto fundamental, ya que puede influir significativamente en la fiabilidad y validez de los datos obtenidos. Para estructurar esta secuenciación, se pueden adoptar dos estrategias principales:

- Codificación inicial con *ChatGPT*. En esta estrategia, el modelo de inteligencia artificial genera una primera codificación de la totalidad del contenido. Posteriormente, codificadores humanos independientes (generalmente uno) o herramientas automatizadas adicionales, revisan y codifican el conjunto completo de datos o un porcentaje (a poder ser representativo y aleatorio) de las inferencias realizadas por *ChatGPT*. Cuanto mayor sea el porcentaje de revisión por parte de los codificadores humanos o de otra plataforma, mayor será la certeza sobre la fiabilidad del análisis de contenido. Esta estrategia es ampliamente recomendada por razones prácticas, ya que permite delegar la mayor carga de trabajo inferencial a *ChatGPT*. Dado que *ChatGPT* es quien realiza la codificación inicial y en su totalidad, su codificación es la utilizada para el análisis de los resultados si la fiabilidad obtenida en las pruebas estadísticas es alta.
- Codificación inicial con humanos. En esta estrategia, un codificador humano independiente realiza la codificación completa de todas las observaciones antes de que *ChatGPT* procese el mismo conjunto de datos o un subconjunto de las observaciones. Dado que la segunda codificación realizada por *ChatGPT* es completamente automatizada, se recomienda aplicar la codificación a la totalidad de las observaciones en lugar de a una fracción, con el fin de evaluar con mayor precisión la fiabilidad del análisis. No obstante, esta secuenciación presenta una desventaja en términos de carga de trabajo, ya que requiere que el codificador humano analice previamente todas las observaciones, lo que puede suponer un esfuerzo considerable en estudios con grandes volúmenes de datos. Dado que es el codificador humano quien realiza la codificación inicial y en su totalidad, su codificación es la utilizada para el análisis de los resultados, siempre y cuando la fiabilidad mostrada en las pruebas estadísticas con sea alta.

Paso 4. Análisis de fiabilidad y validez

El cuarto paso del proceso de análisis consiste en evaluar la fiabilidad y la validez de las inferencias realizadas. Este proceso incluye los siguientes procedimientos:

Análisis de la validez de los datos

Para evaluar la validez de los datos en un análisis de contenido, es fundamental contar con una "verdad de referencia" (*ground truth*). En su ausencia, únicamente es posible examinar la fiabilidad de las inferencias, pero no su validez, una situación frecuente en la mayoría de los análisis de contenido. Cuando el análisis se lleva a cabo con codificadores humanos, la comunidad científica suele confiar más en la validez de los datos, especialmente cuando estos han sido evaluados por codificadores bien entrenados. Sin embargo, cuando las codificaciones son generadas por modelos

automatizados, la validez tiende a ser evaluada con mayor escepticismo. Un ejemplo ilustrativo es el análisis de sentimiento aplicado a reseñas de productos, en el que se clasifican los comentarios de los usuarios como "positivos", "neutros" o "negativos". Si dos sistemas de inteligencia artificial realizan esta tarea de forma independiente y producen resultados idénticos, esto no implica necesariamente que sus inferencias sean correctas: ambos modelos podrían estar cometiendo errores de forma consistente. Desafortunadamente, la única manera de verificar la validez en estos casos es:

- Contar con una "verdad de referencia" para contrastar los resultados obtenidos.
- Utilizar codificadores humanos independientes entrenados, ya que su juicio es generalmente considerado como el estándar de referencia más confiable.

En relación con las estrategias de secuenciación de codificadores discutidas anteriormente, cuando toda la codificación es realizada inicialmente por *ChatGPT* y posteriormente una muestra aleatoria es revisada por un codificador humano, el análisis puede alcanzar una alta fiabilidad. No obstante, la certeza sobre su validez puede depender de la naturaleza del contenido y de los criterios utilizados para evaluar las inferencias, ya que el grueso del proceso sigue dependiendo de la IA. Por otro lado, en la estrategia inversa, donde la codificación inicial es realizada por un ser humano, esta puede considerarse como una posible referencia (*ground truth*) frente a la cual comparar las inferencias generadas por *ChatGPT*. En este caso, se podría asumir un mayor grado de validez si se parte del (discutible) supuesto de que la codificación humana es más precisa que la automatizada mediante *ChatGPT*. Sin embargo, conviene matizar que investigaciones previas centradas en el análisis de contenido latente han documentado que *ChatGPT* puede generar mejores inferencias que los propios codificadores humanos. En cualquier caso y de modo general, cuando se dispone de evidencia sólida sobre la validez de las inferencias (bien sea humana o automatizada), es recomendable complementar los estadísticos de fiabilidad tradicionales con métricas adicionales de validez, las cuales se explican en la siguiente sección.

Examen de la validez con una verdad de referencia

El *accuracy* o precisión global es una métrica que mide la proporción de predicciones correctas en relación con el total de predicciones realizadas. Es una medida ampliamente utilizada para evaluar la validez de un modelo de clasificación, ya que proporciona una visión general de su rendimiento. La precisión se calcula mediante la siguiente fórmula:

$$\text{Precisión global} = \frac{\text{Verdaderos Positivos} + \text{Verdaderos Negativos}}{\text{Verdaderos Positivos} + \text{Verdaderos Negativos} + \text{Falsos Positivos} + \text{Falsos Negativos}}$$

donde:

- Verdaderos Positivos: casos correctamente clasificados como positivos.
- Verdaderos Negativos: casos correctamente clasificados como negativos.
- Falsos Positivos: casos incorrectamente clasificados como positivos.
- Falsos Negativos: casos incorrectamente clasificados como negativos.

Esta es la fórmula de referencia cuando la codificación cuenta con dos clases. Se puede generalizar para *n* clases como la ratio entre el número de aciertos y el número total de casos. Como contamos con la verdad de referencia (*ground truth*), podemos comparar cada método de codificación con el resultado los codificadores humanos obteniendo una precisión global para cada método.

La precisión global varía entre 0 (nula precisión) y 1 (precisión perfecta). Aunque no existe un umbral universalmente aceptado, en la mayoría de los contextos se considera que una precisión superior a 0.9 es excelente; entre 0.7 y 0.9 es buena, pero puede requerir mejoras dependiendo del contexto;

inferior a 0.7 sugiere un rendimiento deficiente necesitando revisión. Es importante tener en cuenta que debe interpretarse en función del contexto del estudio y la naturaleza de los datos. Sin embargo, la precisión puede ser engañosa en casos de clases desbalanceadas, donde una clase es significativamente más frecuente que la otra. En tales casos, es recomendable complementar la precisión con otras métricas como la precisión, la sensibilidad o el F1-score.

El *F1-score* (**Yakouby**; **Axman**, 2020) es una métrica que combina dos medidas clave en la evaluación de modelos de clasificación: la precisión y la sensibilidad. Es especialmente útil en situaciones donde existe un desbalance entre las clases, ya que proporciona un equilibrio entre la capacidad para identificar correctamente los casos positivos (sensibilidad) y para evitar falsos positivos (precisión). Se calcula mediante la siguiente fórmula:

$$F1 - score = 2 * \frac{\text{Precisión} * \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}$$

donde,

$$Precisión = \frac{\text{Verdaderos Positivos}}{\text{Veraderos Positivos} + \text{Falsos Positivos}}$$

Y

$$\text{Sensibilidad} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}}$$

La interpretación y umbral de referencia es análoga a la empleada para la precisión.

Análisis de fiabilidad de la codificación

Diferentes descriptivos y test estadísticos pueden utilizarse para la evaluación de la fiabilidad. Los más destacados son:

- Número de acuerdos: cantidad de casos en los que los codificadores asignan la misma categoría a la misma unidad de análisis.
- Número de desacuerdos: cantidad de casos en los que los codificadores asignan categorías diferentes a la misma unidad de análisis.

$$Porcentaje\ de\ acuerdos = \left(\frac{\text{Número de acuerdos}}{\text{Número total de observaciones}}\right) * 100$$

El porcentaje de acuerdos calcula la frecuencia con la que los codificadores coinciden en un conjunto de juicios en relación con el número total de juicios realizados.

$$Porcentaje\ de\ desacuerdos = \left(\frac{\text{Número de desacuerdos}}{\text{Número total de observaciones}}\right) * 100$$

El porcentaje de desacuerdos calcula la frecuencia con la que los codificadores discrepan en un conjunto de juicios en relación con el número total de juicios realizados.

$$Alpha\ de\ Krippendorff = 1 - \frac{D_o}{D_e}$$

donde, D_o representa la suma de las diferencias al cuadrado entre las frecuencias asignadas por los codificadores para cada ítem. Por otro lado, D_e es el nivel de discrepancia que se esperaría por azar, calculado en función de la distribución de las frecuencias a lo largo de todos los ítems. Para

una mejor compresión de las suposiciones y alcance del estadístico se recomienda leer la literatura especializada (**Hayes**; **Krippendorff**, 2007; **Goyanes**; **Piñeiro-Naval**, 2024). El rango de este estadístico es de 0 a 1, siendo .667 el límite inferior admisible, e igual o superior a .800 lo más apropiado (**Krippendorff**, 2004).

$$Kappa\ de\ Cohen = \frac{P_o - P_e}{1 - P_e}$$

Donde P_o es la proporción de acuerdo observado (es decir, el nivel real de coincidencia entre los codificadores) y P_e es la proporción de acuerdo esperado (es decir, el acuerdo que ocurriría por azar). La **Kappa de Cohen** es una medida estadística utilizada para evaluar el grado de concordancia entre codificadores en datos categóricos. Esta prueba permite determinar en qué medida dos codificadores independientes coinciden en sus juicios, más allá del acuerdo que podría esperarse por azar (**Goyanes**; **Piñeiro-Naval**, 2024). Los rangos propuestos para explicar la fortaleza de la fiabilidad son:

- <0.20 = Fiabilidad entre codificadores pobre
- 0.21 – 0.40 = Fiabilidad entre codificadores discreta
- 0.41 – 0.60 = Fiabilidad entre codificadores moderada
- 0.61 – 0.80 = Fiabilidad entre codificadores sustancial
- 0.81 – 1.00 = Fiabilidad entre codificadores muy buen

Paso 5. Replicabilidad y transparencia

El último paso del protocolo metodológico establece una serie de procedimientos específicos para facilitar la replicabilidad y transparencia en el análisis de contenido. En particular, se proponen las siguientes estrategias:

Publicar el conjunto de datos utilizado en un repositorio abierto

Se recomienda que tanto el contenido analizado como la codificación realizada por los distintos agentes involucrados se hagan accesibles al público para garantizar la replicabilidad y transparencia del estudio. Para ello, es conveniente utilizar plataformas especializadas en el almacenamiento y difusión de datos científicos, como *OSF* (*Open Science Framework*) o *GitHub*, que permiten compartir conjuntos de datos, scripts y documentación de manera estructurada y accesible para la comunidad investigadora.

Descripción de cualquier intervención realizada después del análisis

Es fundamental documentar cualquier intervención manual o automatizada realizada durante el proceso de análisis. Aunque *ChatGPT* permite la codificación automatizada, los investigadores pueden necesitar realizar ajustes, correcciones o verificaciones adicionales tanto en los datos generados por la IA como en los codificados por humanos, ya sea por la detección de errores o por una fiabilidad insuficiente en el análisis. En este sentido, resulta esencial reportar todas las modificaciones realizadas, incluyendo reformulaciones de *prompts*, revisiones de categorías y ajustes en la codificación, con el objetivo de garantizar la transparencia y la replicabilidad del estudio.

Lista de verificación final

A continuación, en la tabla 1, se presenta la lista de verificación que los investigadores pueden utilizar como referencia para garantizar la rigurosidad del procedimiento empleado en el análisis de contenido asistido por *ChatGPT*.

Tabla 1. Lista de verificación para la realización de análisis de contenido asistido por *ChatGPT*

Paso 1. Documentación del Proceso de análisis con *ChatGPT*	¿Se han registrado las instrucciones utilizadas para el análisis de contenido con *ChatGPT*?
	¿Se ha especificado la versión de *ChatGPT* utilizada?
	Si hubo preprocesamiento de los datos procedentes del contenido, ¿se ha descrito convenientemente?
Paso 2. Explicación del libro de códigos y su aplicación en la inferencia	¿Se han definido las variables y el libro de códigos utilizado?
	¿Se han establecido reglas claras sobre cómo codificar cada observación?
	¿Hay coherencia entre el entrenamiento y las reglas del codificador humano y las instrucciones a ChatGPT?
Paso 3. Análisis de consistencia y secuenciación de la codificación	¿Se ha evaluado la consistencia interna de *ChatGPT*?
	¿Se ha establecido la secuenciación del proceso de codificación?
Paso 4. Análisis de fiabilidad y validez	¿El análisis de contenido cuenta con una verdad de referencia?
	Si los datos cuentan con una verdad de referencia, ¿se ha ejecutado una prueba estadística como la precisión o el F1-score?
	¿Se han reportado estadísticos descriptivos y pruebas que evalúen la validez de la codificación?
Paso 5. Replicabilidad y transparencia	¿Se ha publicado en un repositorio en abierto el contenido y la codificación efectuada?
	Si se ha efectuado alguna intervención después del análisis, ¿se ha reportado y descrito convenientemente?

Consideraciones finales

Conviene destacar, finalmente, que los investigadores deben asegurarse de que los datos puestos a disposición pública estén anonimizados (en caso de emplear material de personas físicas) y cumplan con la normativa vigente en materia de protección de datos. Asimismo, es necesario que la comunidad científica reflexione sobre cómo los modelos de inteligencia artificial pueden reflejar o amplificar sesgos presentes en los datos de entrenamiento. En este sentido, resulta fundamental que los distintos campos científicos afectados por la inteligencia artificial se esfuercen por identificar dichos sesgos y adopten medidas para mitigarlos cuando sean detectados. Por último, aunque la inteligencia artificial permite automatizar gran parte del análisis de contenido, la supervisión humana sigue siendo esencial. En última instancia, son los humanos los que asumen la responsabilidad final de las inferencias realizadas.

Referencias

Goyanes, Manuel; **De-Marcos, Luis**; **Domínguez-Díaz, Adrián** (2024). "Automatic gender detection: a methodological procedure and recommendations to computationally infer the gender from names with ChatGPT and gender APIs". *Scientometrics*, v. 129, n. 11, pp. 6867-6888.
https://doi.org/10.1007/s11192-024-05149-2

Goyanes, Manuel; **Piñeiro-Naval, Valeriano** (2024). "Análisis de contenido en SPSS y KALPHA: Procedimiento para un análisis cuantitativo fiable con la Kappa de Cohen y el Alpha de Krippendorff". *Estudios sobre el mensaje periodístico*, v. 30, n. 1, pp. 123-140.
https://doi.org/10.5209/esmp.92732

Jiang, Qile; Gao, Zhiwei; Karniadakis, George-Em (2025). "DeepSeek vs. ChatGPT vs. Claude: A comparative study for scientific computing and scientific machine learning tasks". Theoretical and Applied Mechanics Letters, v. 15, n. 3.
https://doi.org/10.1016/j.taml.2025.100583

Krippendorff, Klaus (2004). "Reliability in content analysis: Some common misconceptions and recommendations". *Human communication research*, v. 30, n. 3, pp. 411-433.
https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Hayes, Andrew F.; **Krippendorff, Klaus** (2007). "Answering the call for a standard reliability measure for coding data". *Communication methods and measures*, v. 1, n. 1, pp. 77-89.
https://doi.org/10.1080/19312450709336664

Maj, Agnieszka; **Makowska, Marta**; **Sacharczuk, Katarzyna** (2025). "The content analysis used in nursing research and the possibility of including artificial intelligence support: A methodological review". *Applied nursing research*, v. 82, 151919. *https://doi.org/10.1016/j.apnr.2025.151919*

Mervaala, Erkki; **Kousa, Ilona** (2025). "Out of Context! Managing the Limitations of Context Windows in ChatGPT-4o". *Journal of data mining & digital humanities*, jdmdh:15090.
https://doi.org/10.46298/jdmdh.15090

Moon, Hak (2025). Comparación del rendimiento de modelos lingüísticos extensos en problemas de cálculo avanzado. arXiv.
https://doi.org/10.48550/arXiv.2503.03960

Riffe, Daniel; **Lacy, Stephen**; **Fico, Frederick**; **Watson, Brendan** (2019). *Analyzing media messages. Using quantitative content analysis in research* (4th Ed.). Routledge.

Yacouby, Reda; **Axman, Dustin** (2020). Probabilistic extension of precision, recall, and F1 Score for more thorough evaluation of classification models. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems* (pp. 79-91). *https://doi.org/10.18653/v1/2020.eval4nlp-1.9*