Proyecto Espacio de Observación de Inteligencia Artificial en Español. Ámbito 0.2. Diseño y cálculo de la métrica agregada para medir la brecha español/ingles en tecnologías de la lengua. Informe Año 2.

Enrique Amigó[1], Jorge Carrillo-de-Albornoz[1], Andrés Fernández[1], Julio Gonzalo[1], Miguel Lucas[2], Guillermo Marco[1], Roser Morante[1], Jacobo Pedrosa[1], Laura Plaza[1], Eva Sánchez[1], Augusto Villa[2]
[1] Natural Language Processing and Information Retrieval Group, UNED
[2] LLorente & Cuenca Madrid, S.L.
Autor de contacto: Julio Gonzalo - julio@lsi.uned.es

Índice

Resumen ejecutivo

En el marco del proyecto ODESIA (espacio de observación para la Inteligencia Artificial en español, fruto de un convenio entre Red.es y UNED financiado por la Estrategia Nacional de Inteligencia Artificial), se ha realizado una estimación de la brecha de desarrollo de la Inteligencia Artificial en inglés y en español para el Año 2 del proyecto. Como en el Año 1, esta brecha se ha medido en cuatro ámbitos: (1) estado del arte de las tecnologías del lenguaje; (ii) soluciones de mercado; (iii) nivel de adopción de la tecnología; y (iv) experiencia de uso.

En la primera iteración del proyecto se realizó un estudio en profundidad de los dominios y tipos de problemas a nivel abstracto (clasificación, etiquetado, ranking, etc.) en las tecnologías del lenguaje con el fin de asegurar una buena cobertura en el estudio de la brecha lingüística. A lo largo del último año las tecnologías basadas en grandes modelos de lenguaje ha revolucionado la capacidad de los sistemas de resolver problemas diversos. Por ello, se han revisados las tipologías de tareas y las dimensiones de evaluación de sistemas inteligentes en el contexto de las tecnologías del lenguaje, incluyendo aspectos como los sesgos en las respuestas, contenidos no informativos o engañosos, competencias cognitivas de los sistemas, etc. El análisis de estas dimensiones se ha reflejado tanto en el diseño de nuevos datasets como en la elaboración de encuestas.

Los resultados, que pueden verse como tabla resumida en la Figura 1, son los siguientes:

- **Ámbito 1 (Estado del arte): brecha global del 66 %**. La brecha promedio sobre todos los aspectos medidos es similar a la del año 1. Dentro del estado del arte, en cuanto a diseminación y recursos, se mantiene la tendencia observada en la primera iteración del proyecto, con algunas diferencias. Al igual que en la iteración anterior, el factor más desfavorable es la diseminación, con una **brecha en publicaciones y proyectos subvencionados del 98 % y 96 % respectivamente**. En concreto, la brecha en proyectos subvencionados ha ascendido del 88 % al 96 % respecto del año anterior. En cuanto a recursos, **la disponibilidad de textos en internet (R.0) se mantiene estable**, como era de esperar dado que no es un indicador susceptible de cambios bruscos. **La brecha en disponibilidad de modelos de lenguaje se mantiene también muy similar (R.1)**. Se observa un **aumento importante de la brecha en disponibilidad de datos anotados en repositorios (R.2)**, sobre todo debido al incremento de datos para el inglés en Hugging Face. La presencia de datos de campañas de evaluación, sin embargo, permanece bastante constante en las fuentes consideradas, si tenemos en cuenta el reducido número de muestras y el consiguiente efecto en la volatilidad del indicador R.2.b. El mayor esfuerzo de medición en este ámbito ha sido para calcular la brecha de efectividad de los modelos de lenguaje:
  - En el segundo año, **hemos pasado de 6 a 10 tareas discriminativas en el leaderboard ODESIA CORE** (con datos generados en el proyecto), para un total de 15 tareas discriminativas en el leaderboard ODESIA EXTENDED (que incluye 5 datasets más de dominio público). En cuanto a tareas abstractas se refiere, para la estimación de la brecha, se han cubierto la clasificación binaria (EXIST 2022 tarea 1, EXIST 2023 tarea 1, DIPROMATS 2023 tarea 1), la clasificación multiclase, jerárquica y/o multilabel (EXIST 2022 tarea 2, EXIST 2023 tareas 2 y 3, DIPROMATS 2023 tareas 2 y 3), la evaluación en modo *learning with disagreement* (EXIST 2023, tareas 1,2 y 3), la regresión (STS 2017), y el etiquetado de secuencias (DIANN 1 y 2). Dentro de lo que se consideran problemas dinámicos, en la estimación de la brecha consideramos el *question answering* con anotación de secuencias (SQUAD/SQAC 2024). Además, hemos completado dos datasets adicionales para medir la efectividad de modelos generativos: UNED-ACCESO (de exámenes tipo test de once asignaturas de acceso a la universidad) y CURIA (de resúmenes en lenguaje claro de textos jurídicos).
  - En conjunto, los dominios y áreas de aplicación cubiertos en el segundo año en el cálculo de la brecha en efectividad incluyen: geopolítica y desinformación (DIPROMATS), biomedicina y extracción de información (DIANN), publicaciones académicas y machine reading (SQUAD/SQAC), redes sociales y contenidos tóxicos (EXIST), noticias (MLDOC), conocimiento enciclopédico y consultas en buscadores (MULTICONER), resolución de similitud

Figura 1: Estimación ODESIA de la brecha entre español e inglés, Año 2

textual (STS), información jurídica y resúmenes en lenguaje claro (CURIA), y exámenes de conocimiento general (UNED-ACCESO).

- **Sobre las tareas discriminativas se ha medido en el leaderboard ODESIA EXTENDED una brecha promedio del** 20 ± 06 **%, consistente con la medición del año anterior**. Hay que destacar que la brecha es positiva en todas las tareas discriminativas evaluadas, excepto en una. Es decir, independientemente del problema abordado, los modelos tienen una efectividad menor en español que en ingles para tareas equivalentes. Otro resultado destacable de este estudio es que **los modelos de lenguaje en español no obtienen mejores resultados que los modelos multilingües equivalentes en español**.
- Más allá de las actividades previstas en el convenio, ante la irrupción de la IA generativa se ha comenzado a evaluar en ODESIA la brecha de rendimiento de modelos generativos. Se han realizado dos experimentos: (i) se ha evaluado GPT-4 en modo zero-shot sobre tres tareas discriminativas del leaderboard, en las que se ha obtenido una **brecha promedio del 18 % en el rendimiento de GPT-4 en inglés y español**. Esta cifra es compatible con la brecha medida para los modelos discriminativos, aunque hay que ampliar la experimentación para consolidarla e incluirla en la medición de la brecha; y (ii) se han evaluado seis modelos generativos (GPT-4, Claude 3 Opus, GPT-3.5, Llama-2, Mistral y Gemma) sobre el dataset de exámenes UNED-ACCESO desarrollado dentro del proyecto. En este caso se ha observado una **brecha promedio del 12 % en los modelos abiertos** y ligeramente negativa en los propietarios (-1 %). Esta estimación de la brecha tiene seguramente un sesgo derivado de posible contaminación, ya que las preguntas originales están en español y se han traducido dentro del proyecto. Es decir, es probable que, al menos los modelos propietarios, hayan visto las soluciones a las preguntas de examen en su formato original en español. En conjunto, se requiere más experimentación para medir con fidelidad la brecha de los modelos generativos.

- **Ámbito 2 (Soluciones de mercado): 8 %**. Esta brecha corresponde a la brecha de funcionalidades en productos comerciales disponibles en ambos idiomas y ha descendido un punto respecto al año anterior.

- **Ámbito 3 (Nivel de adopción): 61 %**. En cuanto a la brecha en nivel de adopción, aparecen variaciones a nivel de indicador específico, aunque el promedio se mantiene prácticamente constante. En concreto, ascienden las menciones de productos en medios (I.A.3) de un 76 % a un 83 %, las menciones de tecnologías de la lengua en medios de un 49 % a un 70 % (I.A.4) y la brecha en adopción para uso personal (I.A.7) que asciende de un 33 % a un 36 %. Sin embargo, descienden la brecha en menciones de productos informes (I.A.1) de un 95 % a un 93 %, las menciones de tecnologías en informes (I.A.2) de un 56 % a un 53 %, el impacto de las tecnologías en la empresa (I.A.5) de un 66 % a un 60 % y la adopción para uso profesional (I.A.7) de un 46 % a un 33 %.

- **Ámbito 4 (Experiencia de uso): -6 %**. En cuanto al ámbito de experiencia de usuario, también se mantiene constante en promedio, aunque hay variaciones a nivel de indicador específico. Se ha obtenido el mismo patrón que el año anterior. Los indicadores de polaridad reputacional (I.E.1) y curvas de valor (I.E.2) donde los indicadores se estiman a partir de opiniones en la web, aparece una brecha negativa en favor del español. Lo mismo ocurre en el indicador de limitaciones (I.E.4) en donde se encuesta a individuos sobre las deficiencias específicas de los productos analizados. Sin embargo, en el caso de las encuestas de satisfacción (I.E.3) los usuarios de tecnologías en inglés se sienten más satisfechos, efecto que ha crecido en los resultados de este año (12 % frente al 2 % obtenido en el año anterior). Esto se compensa con la reducción de brecha en favor del español en los otros tres indicadores (-9 % frente a 2 %, -9 % frente a -4 % y 16 % frente a 25 %). De manera adicional, se han ampliado las encuestas sobre limitaciones para identificar aspectos de la calidad de las tecnologías definidos en la sección 3 de este documento.

En conjunto, hemos medido de nuevo una brecha significativa en casi todos los ámbitos estudiados, lo

que confirma la necesidad de impulsar la IA en español como parte de cualquier estrategia nacional de desarrollo tecnológico.

1. Introducción

Este documento contiene el diseño y cómputo de la métrica agregada empleada para medir la brecha en tecnologías de la lengua entre el español y el inglés durante el segundo año del del proyecto del Espacio de Observación de Inteligencia Artificial en Español, que se lleva a cabo mediante un convenio entre Red.es y la UNED. En concreto, este trabajo se ha desarrollado dentro del "Ámbito 0, Coordinación, agregación y diseminación de resultados, Actividad 0.2 Diseño y cálculo de la métrica agregada para calcular la brecha inglés–español". Este documento se corresponde con la segunda iteración del proyecto, realizada desde marzo de 2023 hasta marzo de 2024.

En este documento emplearemos la misma terminología que en el informe del año anterior. Definimos de nuevo los conceptos principales para facilitar la lectura del mismo. Llamaremos *tarea* dentro de las tecnologías de la lengua a un mapeo entre un espacio de entrada y un espacio de salida o de acción, donde al menos uno de ellos contiene expresiones en lenguaje natural (Schlangen, 2021). Esto incluye, por ejemplo, clasificación, generación de texto, extracción de información y sistemas de búsqueda, además de tareas puramente lingüísticas como lematización, análisis sintáctico o reconocimiento de entidades nombradas. Estas tareas representan el núcleo de las aplicaciones de tecnologías de la lengua. Consideraremos como *aplicación* a cualquier programa informático diseñado como una herramienta para realizar operaciones o funciones específicas en el campo de las tecnologías de la lengua. Estas aplicaciones abordan, por tanto, problemas en escenarios específicos, como por ejemplo, clasificación de correos electrónicos en una empresa, traductores automáticos especializados en un dominio, o sistemas de indexación automática de documentos clínicos. Usaremos el término *dominio* para referirnos a un área específica o campo de aplicación. Consideraremos que hay tres tipos de dominios, los específicos (i.e., médico, legal, docente o periodístico), el general y el transversal. Una aplicación es de *dominio específico* si está específicamente diseñada para un dominio de forma que la calidad del comportamiento de la aplicación no es extrapolable a otros dominios. Por otro lado, consideraremos que una aplicación es de *dominio transversal*, si puede ser adaptada para su implementación en diferentes dominios específicos, de forma que la efectividad medida en un dominio puede ser extrapolada, al menos hasta cierto punto, a otros dominios. Por ejemplo, podemos esperar que un asistente conversacional efectivo en el dominio legal pueda ser adaptado a otro dominio con una efectividad similar. Finalmente, consideraremos como *aplicaciones de dominio general* a aquellas aplicaciones que no requieren ningún tipo de adaptación para ser explotadas en diferentes dominios. Algunos ejemplos son la traducción automática (Google Translator, DeepL), los motores de búsqueda en la web (Google Search), los correctores ortográficos (Microsoft Office) o los transcriptores. Para poder analizar las tecnologías de la lengua desde una perspectiva global, hablaremos de *área de aplicaciones* para referirnos a conjuntos de aplicaciones de dominio específico, transversal o general, que procesan datos similares y que contribuyen en escenarios de alguna manera conectados. Por ejemplo, consideraremos que el conjunto de aplicaciones de extracción de información (tarea dentro del procesamiento del lenguaje) dentro del dominio médico conforman una área de aplicaciones dado que aplican metodologías similares a tareas y textos similares, y se explotan en escenarios interconectados.

Siguiendo el mismo esquema que en la iteración anterior, la brecha entre el español y el inglés se analiza para los siguientes ámbitos: (1) dentro del *Ámbito 1: estado del arte*, cercano al mundo académico, estudiaremos la cantidad de trabajos de diseminación y proyectos subvencionados, recursos existentes (corpus de texto, modelos de lenguaje, y datos anotados) y la efectividad de los modelos. (2) En el *Ámbito 2: soluciones de mercado* nos centraremos en el análisis anual de los dispositivos de consumo más relevantes que utilizan tecnologías de Procesamiento del Lenguaje Natural (PLN), sus funcionalidades y características. (3) Dentro del *Ámbito 3: nivel de adopción* estudiaremos la incorporación de tecnologías de la lengua en el entorno industrial. (4) Finalmente, dentro del *Ámbito 4: experiencia de usuario*, estudiaremos el grado de satisfacción por parte de usuarios finales.

En la iteración anterior del proyecto, se presentó un estudio de las tecnologías de la lengua en base a dos dimensiones. Por un lado, desde una perspectiva industrial, se consideraron las diferentes *áreas de*

aplicaciones y *dominios* donde éstas tecnologías son implementadas. Por otro lado, desde una perspectiva técnica, se analizaron las diferentes *tareas de procesamiento de lenguaje*. Este análisis permitió seleccionar un conjunto representativo de tareas para la medición de la brecha en el Ámbito 1: (Estado del Arte, y un conjunto representativo de áreas de aplicación para el estudio de la brecha en los ámbitos de soluciones de mercado, nivel de adopción y experiencia de usuario.

El análisis realizado en el Año 1 se centró fundamentalmente en medir la efectividad de las tecnología, es decir, su capacidad de dar respuestas correctas y satisfacer a los usuarios. Sin embargo, a lo largo del Año 2 del proyecto se ha producido un importante salto cualitativo en el escenario de las tecnologías de la lengua, debido fundamentalmente a la potencia y capacidad de pre-entrenamiento de modelos de lenguaje neuronales sobre grandes colecciones de documentos (GPT, BERT, etc.). En concreto, en este último año, los modelos generativos conversacionales de propósito general han demostrado ser capaces de abordar con éxito problemas complejos (traducción, resumen, búsqueda de respuestas, generación de código de programación, etc.). Tanto es así, que este salto requiere reconsiderar las dimensiones que usamos para evaluar los sistemas de tecnologías de la lengua. Por ello, en esta iteración del proyecto realizamos un análisis en profundidad de los aspectos a tener en cuenta a la hora de evaluar modelos de lenguaje. Algunos de estos aspectos son la generación de contenidos dañinos, la explicabilidad, los sesgos, las competencias internas del sistema (variación lingüística, razonamiento, etc.), la creatividad o la tendencia a generar contenidos engañosos. Este análisis se usa como base para actualizar las encuestas que se usan para medir la brecha en el ámbito de experiencia de usuario y para seleccionar tareas en el ámbito del estado del arte.

Este documento contiene una descripción tanto de la metodología aplicada para definir y calcular los indicadores, como de los resultados obtenidos al aplicarlos en esta segunda iteración del proyecto. Como en el informe del Año 1, en primer lugar se presentan los conceptos básicos (indicadores, dominios y tareas). En segundo lugar, y como novedad respecto al Año 1, se presenta un análisis previo de las dimensiones de evaluación, reflejando las novedades en evaluación en las tecnologías de la lengua. En tercer lugar, se formalizan los indicadores y se define la metodología general aplicada para recopilar la información necesaria para estimar cada uno de ellos. En tercer lugar, se presentan los resultados obtenidos durante la segunda anualidad del proyecto.

2. Indicadores, dominios, y tareas

La Figura 2 muestra un esquema de los indicadores empleados en el estudio de la brecha lingüística durante el primer año del proyecto ODESIA, además de los dominios de aplicación y las categorías de tareas desde un punto de vista técnico. En esta segunda iteración, se mantienen los mismos indicadores, incorporándose nuevas tareas en el Ámbito del Estado del arte y nuevos productos en los Ámbitos de Soluciones de mercado, Nivel de adopción y Experiencia de usuario. Además, se han ampliado las encuestas realizadas para obtener datos en el Ámbito de Experiencia de usuario. Con el fin de que este informe sea auto-contenido, se ha mantenido la descripción detallada de los indicadores presente en el informe del Año 1.

Como consecuencia del salto cualitativo provocado por la aparición de los modelos de lenguaje neuronales en este último año, se hace necesario revisar el análisis tanto de los dominios como de tipos de tareas desde un punto de vista técnico.

2.1. Indicadores

Como muestra la Figura 2, los indicadores se agrupan en dos grandes conjuntos: el de *investigación y desarrollo*, y el de *implantación*. La categoría de indicadores de investigación y desarrollo se estudia en el *Ámbito 1: Estado del arte*. El objetivo es identificar la brecha entre lenguas en cuanto a tecnologías se refiere, independientemente de la implantación de las mismas en la industria. Esta categoría se divide, a su vez, en las siguientes tres subcategorías:

- Indicadores de *diseminación*. Reflejarán la brecha en términos de número de publicaciones y proyectos subvencionados que desarrollan soluciones para las respectivas lenguas.

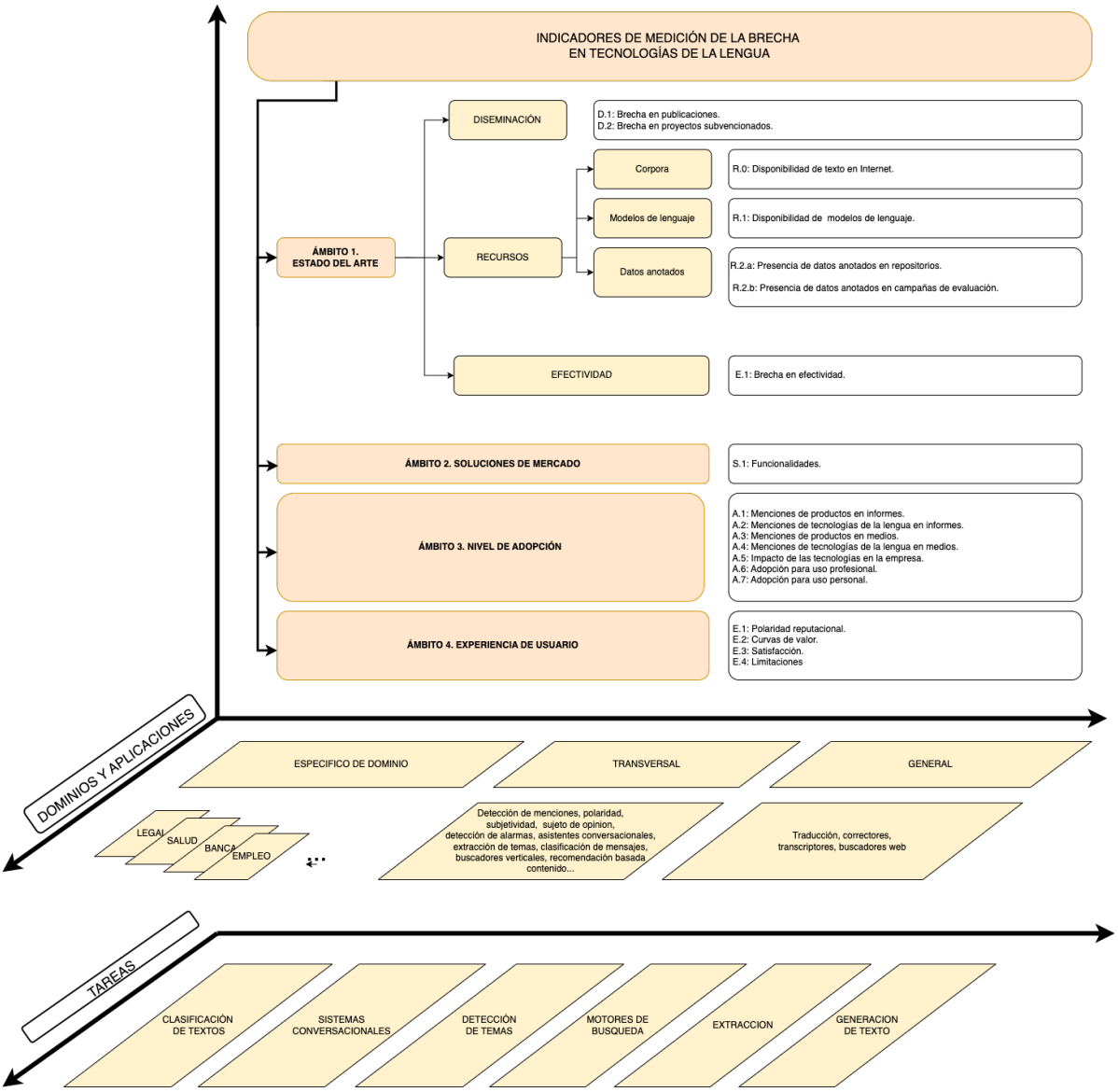

Figura 2: Indicadores para la medición de la brecha en tecnologías del lenguaje agrupados por tipos, dominios de aplicación y tareas en el año 1.

- Indicadores de *recursos*. Analizan la disponibilidad de corpora (textos no anotados sobre los que pre-entrenar los modelos de lenguaje basados en redes neuronales) y de datos anotados sobre los que entrenar y evaluar los sistemas en escenarios concretos.
- Indicadores de *efectividad* en escenarios de uso. Miden la calidad de la salida de aplicaciones orientadas a escenarios de dominio específico, transversal o general. Dentro de esta subcategoría se emplearán dos tipos de fuentes. En primer lugar, estudios recientes de la literatura donde se comparen sistemas en ambas lenguas con datos que sean, hasta cierto punto, comparables. Dado que no existe demasiada literatura en la que los resultados sean directamente comparables, se realizarán experimentos en laboratorio dentro del proyecto comparando la efectividad de modelos neuronales pre-entrenados y modelos entrenados para diferentes tareas.

La segunda gran categoría de indicadores es la de *implantación* de tecnologías, que tendrán como

objetivo cuantificar en qué medida las tecnologías de la lengua son consumidas en el mercado, ofreciendo un valor añadido tanto a empresas como a usuarios finales. Estos indicadores se obtendrán a partir de análisis de aplicaciones, informes de resultados aportados por empresas e instituciones y estudios de opinión en redes sociales. La implantación se evaluará desde los tres ámbitos:

- El *Ámbito 2: soluciones de mercado* tiene como objetivo analizar la oferta de productos que emplean tecnologías de procesamiento en el mercado. Se compararán las funcionalidades que ofrecen diversos productos en inglés y en español.
- El *Ámbito 3: nivel de adopción* tiene como objetivo la estimación del alcance e impacto de la implantación de las tecnologías del lenguaje en español/inglés en empresas. Se estudiará el grado en el que las tecnologías se han incorporado en la industria. Esto incluirá analizar menciones de productos e iniciativas de PLN en informes y medios de comunicación, así como un estudio de las tecnologías adoptadas. Para ello, se emplean encuestas de adopción empresarial y ciudadana.
- El *Ámbito 4: experiencia de usuario* se centra en medir la brecha en el grado de satisfacción de la población en relación a las tecnologías disponibles. Se analizarán las diferencias en base a la polaridad reputacional y curvas de valor en redes sociales.

2.2. Dominios, tareas y áreas de aplicación

Para la iteración del Año 2, en los ámbitos de Soluciones de mercado, Nivel de adopción y Experiencia de usuario se mantienen las áreas de aplicaciones seleccionadas en la iteración anterior. Como área de dominio específico, se ha elegido la **monitorización de reputación on-line** (que incluye análisis del sentimiento), uno de los dominios de aplicación de más impacto en la industria actual dentro de las tecnologías del lenguaje. Como dominio transversal, se han escogido los **asistentes virtuales**. Ésta es también una de las áreas de aplicaciones relacionadas con las tecnologías de la lengua que tiene más impacto en la industria en la actualidad. Finalmente, se han considerado tres áreas de aplicación de dominio general: la **traducción automática**, los **teclados predictivos** y los **buscadores Web** (estos últimos constituyen la aplicación de uso más común entre los ciudadanos).

Incorporación de empresas y soluciones en los ámbitos de soluciones de mercado, implantación y experiencia de usuario en el Año 2 de ODESIA

En el segundo año, se han actualizado los listados de empresas a analizar tanto en Estados Unidos como en España, utilizando el mismo criterio que el año anterior. Por otro lado, en cuanto a las áreas de aplicación concretas, se han incorporado nuevas herramientas a las mismas como ChatGPT, Google Bard y Perplexity. Del mismo modo, se han eliminado otras soluciones utilizando los mismos criterios de selección que el año pasado.

En cuanto al ámbito del estado del arte, y al igual que el año anterior, tanto el estudio de recursos como el desarrollo experimental se ha centrado en evaluar la efectividad de modelos de lenguaje en diferentes problemas. Esta línea de trabajo está en consonancia con el protagonismo actual de los modelos de lenguaje y su posicionamiento como herramienta fundamental en las tecnologías de la lengua. De hecho, se da una tendencia hacia aplicaciones de dominio general y transversal frente a dominio especializado. Más concretamente, existe una tendencia hacia el uso de modelos de lenguaje pre-entrenados a gran escala a los cuales se da instrucciones mediante prompts para que realicen tareas específicas.

Dominios cubiertos en el Ámbito del Estado del arte en el Año 2 de ODESIA

Los dominios cubiertos en el segundo año en el ámbito del estado del arte, para los que se han desarrollado datasets dentro del proyecto, incluyen: geopolítica (DIPROMATS), biomedicina (DIANN), administración y lenguaje claro (CURIA), publicaciones académicas (SQUAD/SQAC), análisis de redes sociales (EXISTS), escritura creativa, y conocimiento general (UNED ACCESO). Además, para la experimentación se han incluido tres datasets adicionales de dominio público que incluyen textos de dominio periodístico (MLDOC), conocimiento enciclopédico y consultas en buscadores (MULTICONER) y resolución de similitud textual (STS).

2.3. Clasificación abstracta de tareas

Las aplicaciones en tecnologías de la lengua se pueden categorizar en función de la tarea que realizan a un nivel abstracto o general, es decir, independientemente del escenario o dominio, como por ejemplo, clasificación de textos, motores de búsqueda, extracción de información, etc. En el informe del primer año del proyecto se agruparon en tres categorías: minería, acceso y generación de texto. Este año consideramos necesario modificar esta categorización debido a la evolución reciente de las tecnologías basadas en modelos de lenguaje. La figura 3 muestra los tipos de tareas en función del output de los sistemas y sus características.

Tareas organizacionales: Se incluye dentro de esta categoría cualquier tarea en la que se clasifique, agrupe, ordene o compare documentos o fragmentos de texto en general. La particularidad es que la respuesta del sistema consiste en textos preexistentes organizados según cierta estructura, a saber, categorías, grupos, ordenación, etc. Dentro de esta categoría, identificamos tareas abstractas como clasificación de textos (ya sean frases, fragmentos o documentos completos), filtrado, ranking o agrupación. Este tipo de tareas abstractas ha sido el foco de atención en las tecnologías de la lengua a lo largo de las últimas décadas, mediante el desarrollo de sistemas de aprendizaje automático supervisados (clasificadores estadísticos), algoritmos de agrupación, y motores de búsqueda. Algunas tareas concretas que entran dentro de esta categoría son la búsqueda de documentos web, análisis de redes sociales, anti-spam, detección de alarmas, etc.

Anotación: Se engloba dentro de tareas de anotación a aquellas tecnologías que identifican estructura *dentro* de los textos. Entran dentro de esta categoría aplicaciones como el análisis sintáctico, árboles de dependencia, desambigüación semántica, reconocimiento y caracterización de entidades nombradas, identificación de roles semánticos, análisis de discurso, etc. También entran dentro de esta categoría aplicaciones como la generación de resúmenes extractivos, el *machine reading*, y múltiples tareas de extracción de información. En un nivel abstracto, podemos distinguir entre anotación de secuencias, donde se incluyen aplicaciones como el reconocimiento de entidades o la búsqueda de respuestas en textos. En un nivel más complejo, tenemos la extracción de estructuras jerárquicas, donde se incluirían aplicaciones como analizadores sintácticos, de discurso, etc.

Generación de lenguaje natural: En estas tareas, la salida consiste en texto generado de forma automática que no está presente de forma literal en los corpus de entrenamiento. Esto incluye sistemas de traducción, transcriptores o sistemas que completan texto. A lo largo del Año 2, han impactado en la sociedad los sistemas conversacionales generativos. Una característica de este tipo de tareas es que permite abordar infinitud de problemas. La razón es que no es necesario convertir la información en los textos en información estructurada, como categorías, entidades o relaciones. En esencia, el objetivo es generar lenguaje a partir de lenguaje. Dado que estas tareas se basan en la previsibilidad del lenguaje, obtener buen rendimiento en estas tareas depende de la disponibilidad de grandes cantidades de textos para crear los modelos de lenguaje de dominios similares al de las tareas a abordar. En un nivel abstracto, podemos distinguir entre tareas en las que se sintetiza o se traduce información textual de entrada y tareas en las que se genera información nueva, como es el caso de los teclados predictivos, los asistentes conversacionales, etc.

Generación de lenguaje formal: Recientemente, la evolución de los modelos generativos ha permitido abordar tareas de generación de lenguaje formal. Esto incluye tanto la traducción de lenguaje natural a proposiciones lógicas (*semantic parsing*), como la generación de código o de órdenes para herramientas (acceso a bases de datos, instrucciones para robots, interacción con interfaces, etc). Este tipo de tareas potencia la comunicación hombre-máquina, traduciendo lenguaje natural a códigos computables. En este caso, el peso no está tanto en las necesidades del usuario, en la información implícita en las fuentes o la previsibilidad del lenguaje, sino en el modelado del problema. Por esta razón, este tipo de tareas está teniendo especial éxito en la generación de código de programación, en donde problemas sencillos como la implementación de un bucle o ciertas funciones han sido modelados en código de programación repetidas veces en las fuentes de entrenamiento. En un nivel abstracto podemos, por nivel de complejidad, distinguir entre generación de instrucciones, tablas, lenguaje lógico-proposicional (proposiciones y operadores lógicos), y código de programación.

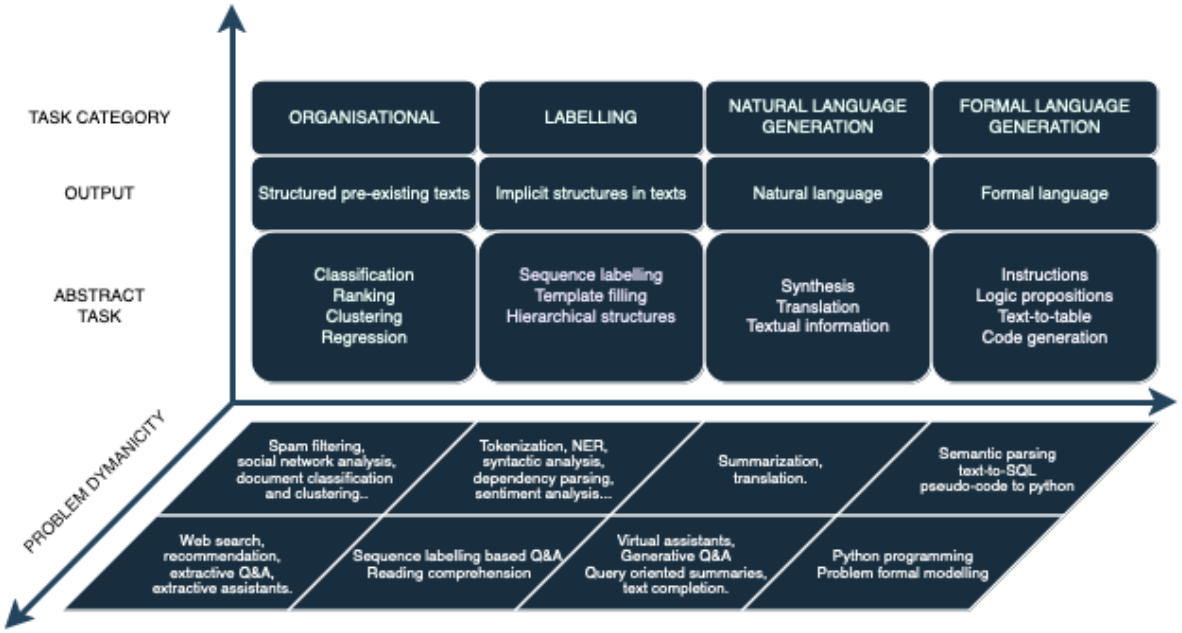

Figura 3: Clasificación de las tecnologías de la lengua en tareas abstractas, con independencia del escenario o dominio.

Por otro lado, desde un punto de vista formal puede distinguirse entre tareas en las que el problema a resolver es estático y aquellas en las que el problema es dinámico. En el primer caso, los criterios de organización, identificación de estructuras o generación de lenguaje no varían. Es decir, en el caso de los problemas estáticos se procesa una colección (organizacional) o texto (anotación) en base a unos criterios fijos. Entrarían dentro de esta categoría, por ejemplo, la clasificación de textos, detección de alarmas, filtros anti-spam, etc. Por otro lado, el problema puede ser dinámico, atendiendo a necesidades específicas en cada caso de test. Este es el escenario de los buscadores web o los problemas de *question answering*, donde el sistema procesa colecciones o textos respondiendo a un problema concreto en cada caso. Por ejemplo, una tarea de preguntas de respuesta múltiple (cuestionarios) se correspondería con una tarea abstracta de clasificación con problema dinámico. Un sistema de resumen automático o de traducción automática sigue las mismas especificaciones cambiando el texto fuente, mientras que un sistema de resumen automático orientado a consulta o un asistente conversacional tiene características dinámicas. En el caso de las tareas de generación de lenguaje formal, el *semantic parsing* atiende a unas especificaciones fijas, mientras que la generación de código de programación responde a problemas dinámicos. En cualquier caso, hay que tener en cuenta que la dinamicidad del problema es una variable continua. Por ejemplo, la generación de código puede atender a un problema más o menos dinámico

dependiendo de si la especificación del programa es muy general o se trata de traducción de pseudo-código a código de programación.

La dinamicidad del problema es un aspecto determinante en el desarrollo de sistemas, dado que limita en gran medida la posibilidad de desarrollar un corpus de entrenamiento adaptado al problema específico que se quiere abordar. Así, problemas estáticos como extracción de información, clasificación de textos, o detección de alarmas son más susceptibles de aprendizaje a partir de muestras de entrenamiento. A medida que un problema es más dinámico, es necesario recurrir a procesos de búsqueda de información y modelos de lenguaje pre-entrenados sobre grandes colecciones para obtener buenos resultados.

Tipos de tareas abstractas cubiertas en el Año 2 de ODESIA

En esta segunda iteración del proyecto, en cuanto a tareas abstractas se refiere, se han cubierto las siguientes: clasificación binaria (EXIST 2022 tarea 1, EXIST 2023 tarea 1, DIPROMATS 2023 tarea 1), clasificación multiclase, jerárquica y/o multilabel (EXIST 2022 tarea 2, EXIST 2023 tareas 2 y 3, DIPROMATS 2023 tareas 2 y 3), *learning with disagreement* (EXIST 2023, tareas 1,2 y 3), generación de texto (CURIA con referencias, y experimentación con escritura creativa, sin referencias), regresión (STS 2017), etiquetado de secuencias (DIANN). Como problemas dinámicos, cubrimos en ODESIA dos tipos de *question answering*: *machine reading* con anotación de secuencias (SQUAD/SQAC 2024) y clasificación en preguntas de respuesta múltiple (UNED ACCESO).

3. Análisis de dimensiones de la evaluación

El reciente desarrollo de modelos de lenguaje pre-entrenados y sistemas generativos nos ha creado la necesidad de revisar, desde una perspectiva general, las diferentes dimensiones sobre las que evaluar las tecnologías de la lengua con el fin de asegurar la cobertura del estudio de la brecha lingüística en tecnologías de la lengua. Un modelo de lenguaje consiste básicamente en una estimación de la distribución de probabilidad de todas las secuencias de palabras posibles dentro de un dominio. En un modelo generativo de lenguaje, esto permite identificar la palabra más probable dada la secuencia anterior. Por extensión, podemos predecir no solo palabras sino etiquetas de clase, lo que nos permite el desarrollo de clasificadores aplicables a todo tipo de problemas. Los modelos de lenguaje neuronales son redes neuronales con múltiples capas entrenadas para predecir palabras o etiquetas en base a secuencias de palabras de entrada. El reciente auge de los modelos de lenguaje neuronales se debe a que, con la disponibilidad de corpus de texto en formato digital y la potencia de computo disponible, estos modelos pueden pre-entrenarse a gran escala sobre grandes colecciones de texto, adquiriendo una elevada capacidad de predicción y generalización. Concretamente, los modelos de lenguaje neuronales superan la capacidad predictiva de un modelo de lenguaje clásico estimado a base de conteo de palabras y, además, superan en capacidad de generalización a los modelos clásicos de representación vectorial de documentos. Como resultado, **los modelos de lenguaje neuronales se han convertido en la herramienta base de las tecnologías del lenguaje**.

Hoy en día, un modelo pre-entrenado es capaz de resolver problemas complejos mediante unas pocas muestras de entrenamiento, obteniendo buenos resultados en diversas tareas como traducción automática (Wu et al., 2016; Vaswani et al., 2017), generación de respuestas (Wang et al., 2018d; Henaff et al., 2017), o inferencia en lenguaje natural (Devlin et al., 2019; Storks et al., 2019), entre muchos otros. Esto lleva a nuevas preguntas sobre cómo deben evaluarse los sistemas basados en estos modelos, la fiabilidad de las respuestas, su falta o exceso de creatividad, los sesgos del lenguaje generado, etc. Actualmente, no existe una metodología consensuada en la comunidad científica para la evaluación de modelos de lenguaje. Encontramos mucha bibliografía donde se abordan diferentes aspectos pero no desde una perspectiva global. Por ello, uno de los objetivos de este informe es **definir una metodología que ofrezca visión global de las posibles dimensiones de evaluación de estos sistemas**. A partir de este análisis, revisaremos los indicadores de brecha lingüística aplicados en ODESIA y su cobertura sobre los diferentes aspectos de calidad de las tecnologías del lenguaje.

Encontramos en la literatura algunos trabajos sobre evaluación de modelos de lenguaje a nivel general, pero estos análisis se articulan sobre tareas o problemas como búsqueda de respuestas, codificación de programas, etc. (Lenci et al., 2021; Chang et al., 2023). Es decir, la calidad de los modelos de lenguaje se mide en función de su efectividad en diferentes tareas o escenarios. Estas tareas son representadas por conjuntos de datos de test consistentes con muestras de entrenamiento, entradas y salidas esperadas. Algunos autores han agrupado estos conjuntos de test en categorías. Por ejemplo, en Guo et al. (2023) se organizan los conjuntos de test existentes en las categorías *knowledge and capability evaluation*, *alignment evaluation* y *safety evaluation*. Sin embargo, en realidad una misma tarea o un mismo conjunto de datos captura o afecta a diferentes aspectos de la calidad del modelo de lenguaje. Por ejemplo, sobre un mismo conjunto de test de búsqueda de respuestas se pueden evaluar múltiples aspectos como la corrección de las respuestas, la fluidez, el sesgo, la originalidad, etc. En este documento partimos de la base de que todos los aspectos de la calidad de un modelo están en mayor o menor medida implicados en cualquier problema, escenario de uso o conjunto de datos de evaluación. Nuestro objetivo es, por tanto, identificar estos aspectos o dimensiones universales de evaluación y determinar qué metodologías de evaluación son más adecuadas para cada dimensión en función de cada tipo de problemas **para así estudiar la adecuación y cobertura de los indicadores y datos analizados en el proyecto ODESIA.**

En este documento estructuraremos el análisis de la evaluación de modelos de lenguajes pre-entrenados sobre una serie de dimensiones de evaluación organizadas en cuatro grupos (ver figura 4):

Calidad de la respuesta: Esta es la dimensión más común en la evaluación de sistemas y se refiere a la calidad de cada una de las salidas del sistema ante cada entrada. Incluye aspectos como la efectividad, la explicabilidad de la respuesta o la generación de contenidos dañinos.

Competencias: Esta dimensión no se centra en la utilidad de la salida del sistema sino en las capacidades que el sistema debe adquirir para generar estas salidas. Esto incluye, por ejemplo, la capacidad de capturar la variación lingüística, la composición de significados o procesos de razonamiento.

Sesgos: Mientras que la calidad de la respuesta se centra en salidas individuales, el análisis de sesgos cubre las salidas de los sistemas en su conjunto. Esto incluye aspectos como la equidad en el tratamiento de diferentes tipos de usuarios, de entradas o de respuestas.

Informatividad y contenidos engañosos: Esta dimensión se centra en la cantidad de información que aporta el sistema en sus soluciones. Incluye aspectos como la originalidad, creatividad o la efectividad para casos poco frecuentes. También tiene que ver con las respuestas engañosas, en cuanto que los sistemas se vuelven engañosos en el momento que producen resultados esperables (de alta probabilidad) pero incorrectos.

En cada una de las siguientes secciones describiremos en detalle estas dimensiones. De forma ortogonal, estructuraremos las tareas en cuatro subconjuntos dependiendo del formato de salida, tal y como se describen en la sección anterior.

En relación a la evaluación de sesgos, el análisis en profundidad de esta dimensión en la evaluación se posterga al siguiente año del proyecto ODESIA. Sin embargo, en esta iteración ya se han incorporado algunos indicadores de sesgo en el ámbito de experiencia de usuario.

Evaluación de sesgo en ODESIA

En relación al sesgo, en ODESIA se aborda este aspecto en el ámbito de la experiencia de usuario, mediante cuestionarios sobre aplicaciones en tareas de acceso a la información (buscadores), tareas de anotación (sistemas de reputación), y de generación de texto (traductores, asistentes virtuales y teclados predictivos). En concreto, los cuestionarios se centran en los sesgos de la salida del sistema: información reputacional sesgada, preferencias por cierto tipo de páginas en los buscadores, sesgos sistemáticos de traducción, etc.

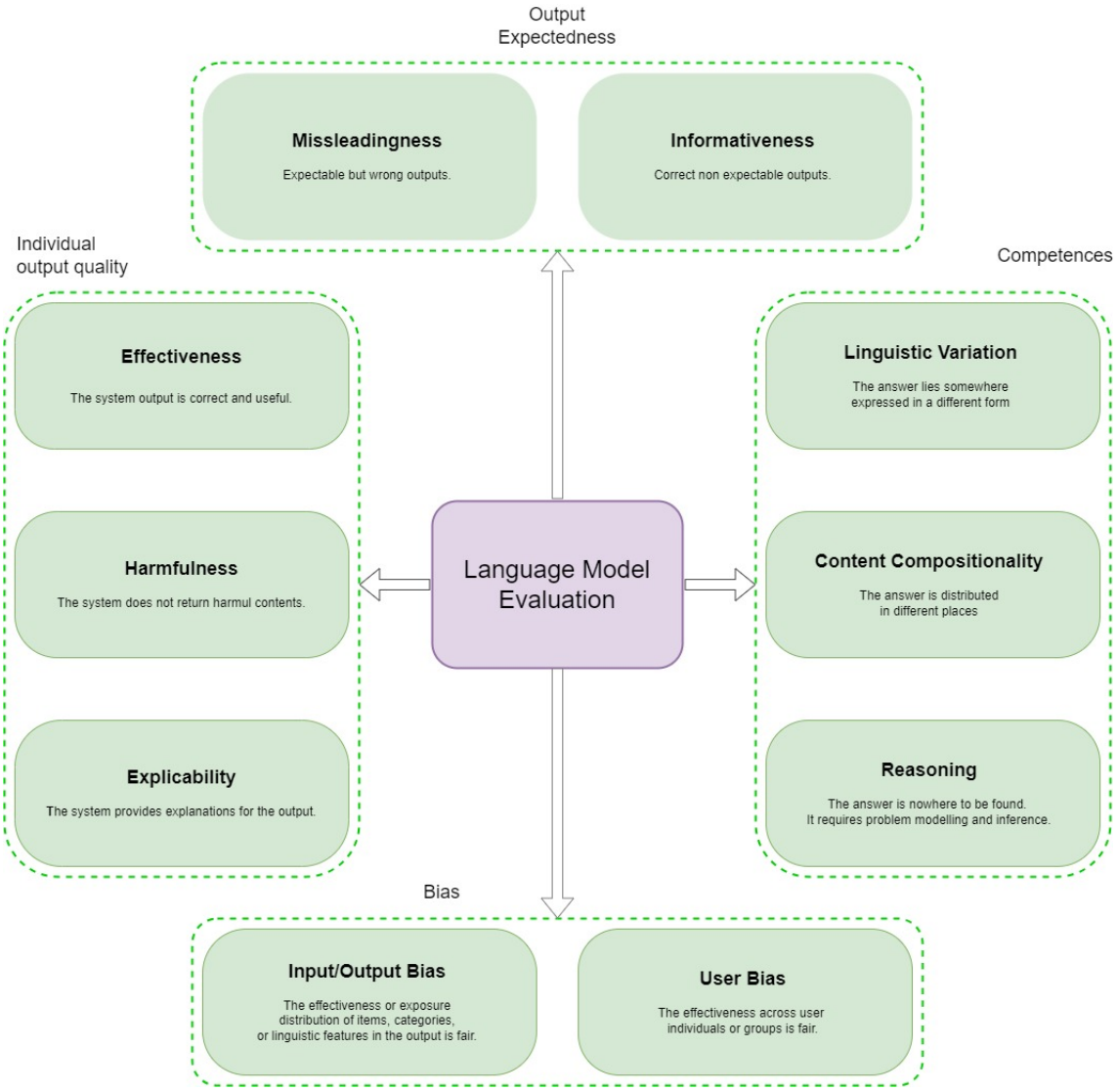

Figura 4: Categorización de dimensiones de evaluación en tecnologías del lenguaje.

3.1. Calidad de las respuestas individuales

Tradicionalmente, la evaluación de la efectividad de los sistemas de inteligencia artificial se ha centrado en valorar la validez o utilidad de cada una de las respuestas del sistema ante cada entrada, ya sea clasificación de elementos de una colección o acierto sobre un conjunto de alternativas para cada entrada, acceso a la información, etc. Dentro de este grupo podemos identificar tres aspectos:

1. La efectividad entendida como la capacidad del sistema de devolver respuestas correctas.
2. La adecuación de contenidos, evitando, por ejemplo, la generación textos dañinos, xenófobos, etc.
3. La capacidad del sistema de, no solo devolver una respuesta correcta, sino de ofrecer una explicación que justifique dicha respuesta.

3.1.1. Efectividad

La efectividad es el criterio de evaluación más común y se refiere a la capacidad del sistema de ofrecer respuestas correctas. La efectividad puede medirse mediante asesores humanos que juzgan la calidad de la salida del sistema en cada caso de test. Esta metodología es cara y no reutilizable. La alternativa es la aplicación de métricas de evaluación que comparan la salida del sistema con una referencia o salida modelo (*gold standard*). En el caso de las métricas de efectividad, existen varios motivos por los que no siempre se emplea la métrica de evaluación más apropiada.

Comparabilidad de resultados: A la hora de elegir una métrica apropiada, en la mayoría de los casos los investigadores optan directamente por la métrica más común. Esto tiene sus ventajas, dado que permite comparar resultados con el estado del arte sobre los mismos datos sin necesidad de repetir experimentos. Por este motivo, las primeras métricas suelen ser las más utilizadas, y no es común que la comunidad cambie de métrica, a no ser que se identifiquen limitaciones muy claras en la métrica original. Por ejemplo, la tasa de aciertos, la precisión y la cobertura siguen siendo las métricas más comunes en clasificación desde hace décadas. La métrica ROUGE (Lin, 2004) para evaluación de resúmenes fue de las primeras que se propuso y sigue siendo, después de 20 años, la más empleada a pesar de sus reconocidas limitaciones. Es muy común que la métrica estándar para un problema sea la que los organizadores de la primera campaña de evaluación definieron y que los organizadores de las campañas subsiguientes la adopten.

Interpretabilidad: El desconocimiento de las propiedades formales y del comportamiento de las métricas hace que los investigadores escojan métricas sencillas de interpretar. Este es el caso de la tasa de aciertos en sistemas de clasificación o de la precisión a profundidad 10 en el caso de los sistemas de recuperación de información. Estas métricas no siempre son las más adecuadas, pero aseguran al investigador y al lector la interpretación de los resultados.

Ausencia de herramientas de evaluación: Existen múltiples plataformas de desarrollo que incluyen una implementación de métricas de evaluación. Por ejemplo, el entorno de desarrollo de aprendizaje automático WEKA (Frank et al., 2005) incluye las métricas estándar de clasificación, pero estas métricas están integradas y el investigador no siempre está haciendo uso de la plataforma en la cual se integran las métricas. En otros casos, el conjunto de métricas incluido en la herramienta puede ser limitado. También supone un problema que las métricas que se quieren aplicar no se encuentren en la misma herramienta. Por ejemplo, la detección de sentimientos (positivo, neutro, negativo), puede evaluarse en términos de clasificación, ordenación o predicción de valores de polaridad. Estas métricas tan dispares no se suelen encontrar en la misma herramienta.

Competición entre grupos de investigación: Si bien las campañas de evaluación como SEMEVAL[1], TREC[2], CLEF[3], etc. surgidas a partir de los 90 han acelerado la investigación mediante la compartición de datos de entrenamiento y test, también han provocado que los grupos de investigación

[1] https://semeval.github.io/
[2] https://trec.nist.gov/
[3] https://www.clef-initiative.eu/

se centren en obtener una buena posición en el ranking de resultados. Una buena posición en una campaña de evaluación supone una ayuda importante a la hora de publicar en revistas o congresos de impacto, solicitar proyectos de investigación, etc. Como comentamos anteriormente, el problema de las campañas de evaluación es que centran toda la evaluación en una única métrica, que no siempre es la más adecuada. En muchos casos, los organizadores de las campañas no disponen del tiempo suficiente para hacer un análisis en profundidad de métricas antes de la primera edición, y resulta complicado cambiar de métrica en las siguientes ediciones de la campaña.

Ausencia de una base formal unificada: Encontramos en la literatura análisis formales alternativos y propiedades de métricas para tareas específicas. Sin embargo, además de múltiples métricas también existen múltiples marcos teóricos y conjuntos de propiedades propuestos por diferentes investigadores. Estas propiedades son en muchos casos incompletas o redundantes, aplicando diferentes terminologías para los mismos conceptos. La inexistencia de un marco teórico unificado sobre el que caracterizar las métricas existentes es también un obstáculo para el acuerdo entre investigadores sobre el uso de métricas. Además, no existe que sepamos un marco teórico fundamental que abarque el problema de la evaluación de forma transversal sobre diferentes problemas, tareas y escenarios.

Ausencia de documentación: Encontramos en la literatura un gran número de publicaciones sobre métricas de evaluación en tareas específicas como clasificación, ranking, agrupación, recomendación etc., en el área de procesamiento de lenguaje y recuperación de información. Existen también documentos que revisan, desde una perspectiva global y formal, las métricas en cada uno de estos problemas. Sin embargo, por lo que sabemos, no existe un documento que aglutine el análisis de métricas en diferentes problemas desde una perspectiva formal o fundamental. En el contexto docente, no existe que sepamos un libro base que cubra el problema de la evaluación sobre referencias a un nivel global. Tampoco aparece en los planes de estudio de informática una asignatura centrada en el problema de la evaluación. En general, se trata de una materia tratada siempre de manera parcial, repartida en distintas asignaturas. Esto impide que los investigadores tengan una formación integrada en el problema de la evaluación. Además, si bien existen libros sobre teoría de la medida en física, o psicometría en psicología, no existe un conocimiento integral sobre los fundamentos teóricos de la evaluación en sistemas inteligentes.

A continuación analizamos la evaluación de la efectividad en los diversos tipos de problemas.

Efectividad en tareas organizacionales y de anotación

En tareas organizacionales y de anotación, las salidas del sistema son estructuras simples: categorías (clasificación), la elección de una respuesta de entre varias alternativas, ítems recuperados de una colección y ordenados según relevancia (buscadores, recomendadores, etc.), ítems o documentos agrupados, plantillas en los sistemas de extracción de información o, en algunos casos, predicciones de valores numéricos. En el área de procesamiento de lenguaje es también muy común el etiquetado de secuencias de palabras (*sequence labelling*), consistente en identificar en un texto fragmentos que se ajustan a una categoría como nombres de entidades, expresiones de opinión, justificaciones, etc.

En general, la efectividad de los modelos de lenguaje para tareas de organización y anotación es, en muchos casos, bastante alta incluso sin muestras de entrenamiento. En Bang et al. (2023) se demuestra empíricamente que ChatGPT obtiene una alta efectividad en tareas discriminativas como la selección de la respuesta correcta ante preguntas que requieren sentido común en conjuntos de datos como CommonsenseQA (Talmor et al., 2019), PIQA (Bisk et al., 2020), o Pep-3k (Wang et al., 2018b). Sin embargo, la efectividad de los modelos de lenguaje en tareas de clasificación puede verse reducida en problemas con alto grado de subjetividad, conjunto numeroso de clases o categorías jerárquicas. Un ejemplo reciente es la tarea definida en la campaña DIPROMATS (Moral et al., 2023-09) consistente en la caracterización de propaganda política en 15 subclases. La medida F1 de los sistemas participantes en 2023 no superó el 0.48 para el inglés o el 0.36 en el caso del español. En tareas de recuperación de información como la

recomendación, el uso de modelos de lenguaje pre-entrenados ha supuesto también un salto respecto a la efectividad de los recomendadores tradicionales, aunque sufren aún muchas limitaciones (Fan et al., 2023).

Las métricas en sistemas discriminativos y de acceso a la información son, en esencia, medidas de similitud que comparan la respuesta generada por el sistema con las anotaciones manuales en el conjunto de test. Estas métricas se pueden categorizar en función de la estructura de salida del sistema. Por ejemplo, métricas de clasificación que evalúan la predicción de categorías, métricas clásicas de recuperación de información que evalúan rankings de documentos, métricas que evalúan grupos de documentos o problemas de cuantificación. En general, la métricas más exitosas en sistemas de clasificación y de recuperación de información son sencillas, asegurando cierta interpretabilidad. Este es el caso de la medida de tasa de aciertos en tareas de clasificación, las métricas DCG (Järvelin and Kekäläinen, 2002) o RBP (Moffat and Zobel, 2008) en recuperación de información o las métricas de pureza y pureza inversa en tareas de agrupación (Zhao and Karypis, 2001).

Sin embargo, en algunos casos se requieren métricas más sofisticadas. Por ejemplo, la incorporación de la teoría de la información en la evaluación lleva a métricas más complejas. Esto lo encontramos en el caso de métricas de clasificación (Amigo and Delgado, 2022), agrupación de documentos (Meilă, 2007) y recuperación de información (Amigó et al., 2022). En el caso de las métricas de acceso a la información, estas pueden sofisticarse al considerar modelos de comportamiento del usuario más complejos (Chapelle et al., 2009; Moffat and Zobel, 2008). Otro de los motivos para el uso de métricas sofisticadas es la necesidad de considerar problemas mixtos, como por ejemplo el ranking con diversificación, donde se persigue maximizar la diversidad de los documentos recuperados (Amigó et al., 2018), la jerarquización de clases en problemas de clasificación (Amigo and Delgado, 2022), o el tratamiento del desacuerdo en los conjuntos de test (Basile et al., 2021).

Efectividad en generación de lenguaje formal

Con el desarrollo de modelos de lenguaje, surge la posibilidad de generar de forma efectiva respuestas en formatos estructurados, es decir, en lenguaje formal. Kamath and Das distinguen, a nivel abstracto, entre tres lenguajes de representación: formalismos lógicos, formalismos basados en grafos y lenguajes de programación (Kamath and Das, 2018). Guo et al. incluyen este tipo de tareas dentro de una categoría más general a la que denominan *tool manipulation* (Guo et al., 2023). Esto se debe a que la generación de lenguaje formal permite la interacción con herramientas que reciben de entrada órdenes o información estructurada. Sin embargo, esta categoría incluye también formatos de salida semi-estructurados (combinación de texto y lenguaje estructurado) como llamadas a buscadores o instrucciones en venta online.

Desde el punto de vista de la evaluación, la manera más natural de categorizar estos sistemas es en base al lenguaje formal generado. Una primera categoría es la traducción de lenguaje natural a formalismos lógicos, lo que comúnmente se denomina *semantic parsing*. Algunos corpus de evaluación para esta tarea son COGS (Kim and Linzen, 2020), GrailQA (Gu et al., 2020), WebQSP (Yih et al., 2016) o FOLIO (Han et al., 2022). Una segunda categoría es la generación de lenguaje de programación, como Python (Austin et al., 2021a; Chen et al., 2021) o SQL (Yu et al., 2018; Xie et al., 2022). También encontramos en la literatura la generación automática de instrucciones de control en robótica, como es el caso del corpus ALFRED (Shridhar et al., 2019).

La evaluación de lenguaje formal se ha tratado desde dos perspectivas: similitud entre la salida del sistema y la referencia o *matching-based*, y equivalencia formal o *functional correctness* (Chen et al., 2021). En la primera aproximación se evalúa la similitud a nivel de lenguaje entre la salida del sistema y la referencia. Es decir, en qué medida se ajusta la secuencia de *tokens* o símbolos generados por el sistema a la secuencia anotada como referencia, sin entrar en la semántica o en la corrección de la salida del sistema cuando ésta es ejecutada. Para ello, en varios estudios se han empleado métricas tradicionales de evaluación de generación de texto, como BLEU, originalmente diseñadas para problemas como resumen o traducción automática. Estas métricas han sido usadas en evaluación de lenguajes formales como Python (Austin et al., 2021b), o SQL (Yu et al., 2019). Sin embargo, este criterio presenta limitaciones. Por ejemplo, para el caso de la generación de código, Ren et al. (2020) identificaron problemas con BLEU a

la hora de capturar la semántica del código generado, por lo que propusieron ciertas modificaciones en la métrica para capturar este aspecto. En el caso de tareas de generación de lenguaje formal en las que la salida del sistema no es compleja, se puede evaluar directamente en base al ajuste exacto entre la salida del sistema y la referencia generada por expertos. Esto se ha aplicado, por ejemplo, en la evaluación de formalismos lógicos (Kim and Linzen, 2020; Li et al., 2021a), o en generación de sentencias SQL (Yu et al., 2018). En este marco de evaluación se asume que a más aciertos, es decir, ajustes exactos entre la salida del sistema y la referencia humana, mayor es la proximidad en general entre ambos. Sin embargo, este método solo es aplicable cuando la salida esperada del sistema es lo bastante simple. La efectividad de los sistemas en base a correspondencia exacta varía en función de la complejidad del problema. Mientras que en Kim and Linzen (2020) se obtubieron tasas del 99 %, en Li et al. (2021a) las tasas de acierto rondaban el 65 %. En el caso de la generación de SQL, en (Yu et al., 2018) se obtuvieron tasas no mayores del 43 % en las consultas más sencillas. En 2022, se reportaron resultados sobre ese mismo corpus de entorno al 75 % (Wang et al., 2022). La efectividad de los sistemas decrece sustancialmente cuando se pasa a herramientas de propósito general. Por ejemplo, en Zhuang et al. (2023) se establece un marco de evaluación en el que los sistemas deben generar instrucciones para el uso de 13 herramientas diversas de acceso a la información sobre diferentes fuentes. En el mejor de los casos, para preguntas sencillas, se llegó a una tasa de aciertos del 40 %.

La segunda opción consiste en evaluar la equivalencia formal entre la salida del sistema y la referencia. Por ejemplo, en Yu et al. (2018) se propone evaluar la coincidencia de los resultados al ejecutar el código SQL sobre una base de datos. Sin embargo, los autores advierten que esta metodología, aunque permite capturar salidas equivalentes expresadas de diferente forma, también puede dar lugar a falsos positivos cuando una salida formal no es correcta, pero genera la misma salida cuando se ejecuta en una base de datos determinada. En evaluación de generación de código Python se ha aplicado como métrica el porcentaje de códigos generados que superan un test escrito por anotadores (Chen et al., 2021; Liang et al., 2023; Kulal et al., 2019). De nuevo, la tasa de aciertos depende mucho de los datos y del problema. En (Kulal et al., 2019), para generación de código c++, se reportó una tasa de aciertos entre el 40 % y el 60 %. En generación de código Python (Corpus HumanEval), se ha reportado una tasa del 67 % (Chalkidis et al., 2021). En generación de código Python para control de robots, en (Liang et al., 2023) se reportó entre un 60 % y 80 %.

Una característica importante de la evaluación de lenguaje formal es que se puede modelar de forma objetiva la complejidad del problema. Por ejemplo, en (Yu et al., 2018) se categorizaron los casos de traducción de texto a SQL en tres niveles (*easy medium* y *hard*) en función de la cantidad y variedad de las instrucciones SQL en la referencia.

En general, existen ciertas limitaciones en los modelos actuales para la generación de lenguaje formal. Por ejemplo, los modelos de lenguaje son de facto empleados para programar. Esto resulta especialmente útil dada la inmensa cantidad de librerías disponibles. La popularidad de uso parece indicar que los niveles de efectividad son aceptables. Sin embargo, una observación habitual es que los modelos necesitan una especificación lo más detallada posible para generar código correcto, o bien ser usados para generar componentes de código sencillos. En ambos casos, supone un esfuerzo extra para el usuario programador. Teniendo en cuenta esto, la evaluación podría enfocarse, además de en la calidad del código generado, en la cantidad de información que es necesario aportar al sistema para que genere código correcto. Este planteamiento puede extenderse a lenguajes formales en general, como proposiciones lógicas, instrucciones SQL, etc.

Efectividad en generación de lenguaje natural

En el caso de sistemas que generan texto libre (salida no estructurada), el problema de la evaluación sigue siendo un tema en discusión. Desde la década de los 2000, los sistemas de generación de texto, en particular los sistemas de traducción automática o resumen automático, han venido evaluándose mediante métricas basadas en solapamiento de palabras o secuencias como ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) o METEOR (Banerjee and Lavie, 2005), así como diferentes distancias de edición. Se han propuesto un gran número de métricas en esta línea como son f-score, CIDER, NIST, GTM, HLEPOR,

RIBES, MASI, WER, TER, DICE (Celikyilmaz et al., 2020). La mayor limitación de estas métricas es que no tienen poco nivel de generalización. No en cuenta aspectos como la estructura del lenguaje o la variabilidad léxica. Como ventaja, son hasta cierto punto objetivas, sin sufrir sesgos de la herramienta de evaluación.

En el siguiente nivel de abstracción, se han desarrollado métricas basadas en alineamiento de embeddings léxicos, como es el caso de BertScore (Zhang et al., 2020), entre otros. También cabe destacar la métrica MAUVE, que se obtiene calculando las divergencias de Kullback-Leibler (KL) entre las dos distribuciones de tokens en un espacio de Embedding cuantizado a partir de un modelo de lenguaje (Pillutla et al., 2021). Otras métricas propuestas en esta línea son MEANT 2.0, YISI, WMD o SMD. Éstas permiten un mayor nivel de generalización que las basadas en solapamiento de tokens, aunque pueden sufrir sesgos dados por la representación semántico-distribucional en la que se basan. Otra limitación es que no capturan en detalle aspectos lingüísticos como la fluidez, coherencia etc.

A medida que los sistemas han ido generando textos de mayor calidad, las métricas basadas en tokens y representaciones distribucionales empiezan a ser insuficientes para comparar textos. En un siguiente nivel encontramos el uso herramientas de procesamiento lingüístico, como son los sistemas de similitud semántica textual, implicación textual, detección de paráfrasis o incluso de comprensión (alineamiento de pasajes y preguntas). En esta línea, Zhong et al. proponen una métrica unificada para tareas de generación de lenguaje basada en Question Answering (Zhong et al., 2022). En Celikyilmaz et al. (2020), se revisan en detalle este tipo de métricas para la evaluación de generación de texto. De esta forma, se obtiene una capacidad de generalización aun mayor, pero a costa de la dependencia de errores y sesgos en estas herramientas.

En un siguiente nivel tenemos sistemas específicamente entrenados para el problema de la evaluación de textos. Por ejemplo, en el caso de BARTScore, donde se aplica un fine-tunning sobre un modelo de lenguaje general (Yuan et al., 2021). Existen diferentes trabajos den donde se entrenan métricas para capturar aspectos específicos de la calidad de los textos como son la consistencia en resumen autmático (Kryscinski et al., 2020a; Wang et al., 2020; Cao et al., 2020) o la coherencia en diálogos (Dziri et al., 2022; Huang et al., 2020; Ye et al., 2021).

En un grado aún mayor de automatización del proceso de evaluación, se ha estudiado el uso de sistemas generativos no entrenados para evaluar textos sin referencias humanas. En concreto, Chiang and Lee obtuvieron una alta correspondencia entre ChatGPT y evaluadores humanos evaluando respuestas de sistemas sobre una escala de Likert. Es importante destacar que los aspectos de calidad evaluados en esta experimentación son de carácter muy genérico, como aspectos linguísticos como gramaticalidad, cohesión, fluidez, consistencia semántica, etc. Además, las tareas evaluadas en la experimentación son de carácter lingüístico, en concreto, la redacción libre de una historia o la generación de variantes de expresiones mediante sinónimos (Chiang and Lee, 2023).

Un estudio exhaustivo de métricas de evaluación de texto reveló que existe una variabilidad significativa en la correlación de diferentes métricas con el juicio humano, es decir, que no hay una métrica que obtenga una alta correlación con las evaluaciones humanas, lo que sugiere la necesidad de una combinación de métricas para lograr una evaluación más completa (Fabbri et al., 2021). Además, aunque algunas métricas contemporáneas superan a las tradicionales en determinados escenarios o dominios, no hay una métrica o familia de métricas que resalte claramente sobre las demás en cuatro dominios clave de análisis: coherencia, consistencia, fluidez y relevancia. En cualquier caso, el uso de métricas entrenadas puede resultar en sesgos por el uso de un modelo pre-entrenado similar al modelo evaluado.

He et al. aplicaron diferentes pruebas de estrés métricas de evaluación de texto, incluyendo métricas clásicas basadas en solapamiento de tokens y métricas basadas en modelos generativos. En general, estas pruebas consistieron en modificar textos de manera artificial para comprobar si las métricas son sensibles a ello. Los autores concluyeron que diferentes métricas son sensibles a diferentes errores. Además, demostraron que algunos métricas basadas en modelos pueden tener auto-sesgo cuando los textos son generados por éstos mismos modelos (He et al., 2023).

En general, el uso de métricas de evaluación de texto libre en base a referencias humanas está condicionado principalmente por dos factores:

1. La variabilidad de las respuestas correctas. En tareas de alta subjetividad, como el resumen automático abstractivo, puede no ser suficientemente representativo disponer de uno o varios ejemplos de resúmenes humanos, mientras que en tareas de dominio más cerrado como la generación de código, el resumen extractivo o la búsqueda de respuestas en fuentes, puede ser más viable usar este tipo de métricas.

2. La calidad de los textos evaluados. En tareas en las que los sistemas son capaces de generar respuestas de alta calidad, las métricas pueden reportar la misma similitud entre textos generados automáticamente y referencias gold, que entre las diferentes referencias generadas por distintos asesores.

3. La dinamicidad del aspecto evaluado. Criterios como la cobertura de contenidos o adecuación de textos son dependientes del input y requieren textos de referencia para ser evaluados, mientras que aspectos lingüísticos como la fluidez o gramaticalidad son independientes del input y permiten con mayor facilidad el entrenamiento de métricas de evaluación.

Evaluación de efectividad en ODESIA

El estudio de la brecha correspondiente al primer año del proyecto cubría experimentos con diez tareas discriminativas, de las cuales seis incluían el desarrollo de datasets propios del proyecto (ODESIA LEADEBOARD CORE v1) y cuatro eran datasets de dominio público (ODESIA LEADERBOARD EXTENDED v1). Las dos tareas de detección de sexismo (EXIST 2022) y las tres de detección de propaganda política (DIPROMATS) fueron evaluadas mediante métricas estándar de clasificación (precisión y cobertura). En la tarea de reconocimiento de entidades centrada en la identificación de discapacidades DIANN 2023, se aplicó una métrica estándar de precisión y cobertura sobre spans, al igual que en la tarea de reconocimiento de entidades nombradas complejas (MutiCONER 2022). En clasificación de noticias (MLDoc) se empleó también F1, así como en la tarea de *Question Answering* extractivo SQUAD/SQAC.

En el segundo año del proyecto se incorporan métricas nuevas, algunas fruto de la investigación en el marco del proyecto. En las tareas de detección de propaganda en tuits (DIPROMATS 2023), dado que se trata de tres problemas de clasificación, uno de ellos con 15 categorías, se empleó una métrica de efectividad más sofisticada, que tiene en cuenta la especificidad de las clases y las relaciones jerárquicas (Amigo and Delgado, 2022). En el caso de las nuevas tareas de EXIST fue necesario diseñar nuevas métricas, ya que no existía anteriormente ninguna métrica aplicable a problemas de clasificación jerárquica multiclase y multietiqueta en el paradigma de *learning with disagreement*. Para ello se definió una extensión de la métrica ICM al paradigma de *learning with disagreement*. Además, se amplían las métricas empleadas en tareas de detección de secuencias, como son MutiCONER, SQUAD/SQAC-2024 y DIANN. Por último, se incorpora en esta iteración un problema de generación de lenguaje, en concreto, una tarea de resumen de textos legales (CURIA-2024), en donde se aplicarán diferentes métricas de evaluación de textos. Finalmente, en el dataset de preguntas tipo test se emplea una métrica de *accuracy* corregida en función de la probabilidad de acierto casual (que varía en función del número posible de respuestas).

Por otro lado, en el ámbito de Experiencia de usuario, la evaluación de efectividad se evalúa en ODESIA mediante cuestionarios sobre productos de acceso a la información (buscadores), de anotación (sistemas de reputación), y de generación de texto (traductores, asistentes virtuales y teclados predictivos).

3.1.2. Contenidos Dañinos

Los contenidos dañinos en respuestas de sistemas de tecnologías de la lengua cubren muchos aspectos. Se refiere a textos generados por el sistema o piezas de información recuperadas con contenidos que puedan afectar a los usuarios y a la población en general. El problema de repuestas dañinas afecta por tanto a tareas de acceso a la información y tareas de generación de texto, siendo menos frecuentes en

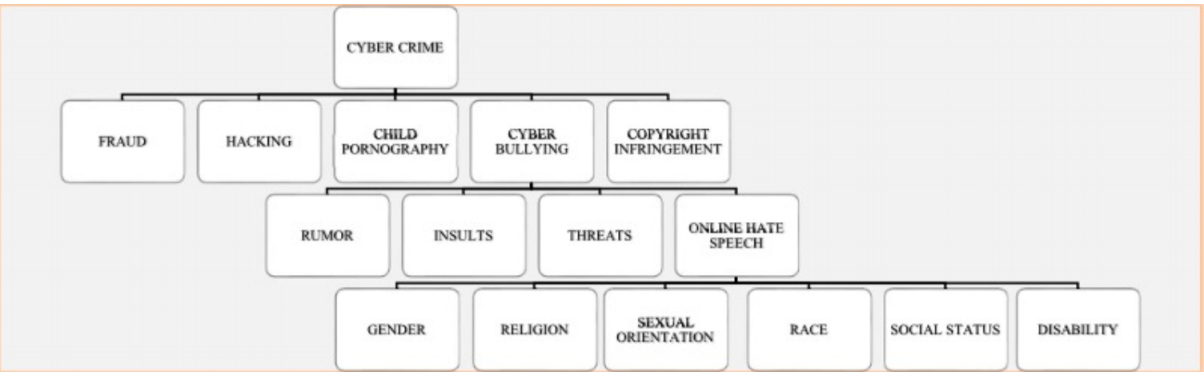

Figura 5: Categorización de tipos de crímenes cibernéticos según Anjum and Katarya (2023).

problemas discriminativos de anotación o de generación de lenguaje formal. Algunas aplicaciones que pueden verse afectadas por este tipo de problemas son los buscadores, recomendadores de contenidos (redes sociales, blogs), o asistentes virtuales generativos. En Anjum and Katarya (2023) se establece una categorización bastante exhaustiva mostrada en la Figura 5 de tipos de crímenes cibernéticos. Dentro de esta taxonomía, la generación de contenidos dañinos se corresponde con la noción de *cyberbullying* y todas sus subcategorías. Estas incluyen rumores falsos que afecten a personas, amenazas, y, en particular, contenidos de odio, dentro de los que se incluyen diferente subcategorías dependiendo del colectivo afectado, que puede ser definido por aspectos como el género, religión, tendencia sexual, raza, estatus social o discapacidades de los individuos que lo conforman.

Desde el punto de vista de la evaluación, identificar contenidos dañinos en la salida de un sistema requiere a su vez un sistema de detección de contenidos dañinos. En otras palabras, las métricas de evaluación de contenidos dañinos son a su vez sistemas de clasificación de textos que requieren una meta-evaluación. A lo largo de estos últimos años, se han desarrollado conjuntos de test y campañas de evaluación para estos tipos de sistemas, como son la tarea de *Hate Speech Detection* en EVALITA (Bosco et al., 2018), racismo y sexismo (Davidson et al., 2017), o lenguaje agresivo (Chatzakou et al., 2017), entre muchos otros.

La efectividad de los sistemas de detección de contenidos dañinos depende en gran medida de la complejidad y subjetividad del problema. Por ejemplo, en la campaña de evaluación EXISTS-2023 para la detección de sexismo se estableció una primera tarea de clasificación binaria (sexista versus no sexista), mientras que otra tarea consistió en clasificar los contenidos sexistas en cinco subcategorías. Los resultados de la evaluación mostraron un salto de efectividad considerable entre ambas tareas.

Evaluación de detección/generación de contenidos dañinos en ODESIA

Tanto en el primer año del proyecto como en el segundo, se han evaluado modelos de lenguaje en la detección de un tipo concreto de contenido dañino, el sexismo. El objetivo en estas dos primeras iteraciones ha sido estudiar la brecha en identificación de contenidos sexista. Además, en el ámbito de Experiencia de usuario, la evaluación de sistemas de detección de contenidos dañinos se realiza en ODESIA mediante cuestionarios sobre aplicaciones, tanto de acceso a la información (buscadores) como de generación de texto (asistentes virtuales).

3.1.3. Explicabilidad

El aspecto más relevante de los modelos de lenguaje es su capacidad predictiva. Sin embargo, una crítica constante a estos sistemas es su naturaleza de caja negra. Es decir, el sistema toma decisiones sin que se pueda establecer cómo ha llegado el sistema a una decisión ni determinar cómo se puede mejorar el sistema para resolver deficiencias específicas. Frente a esto, han surgido en la literatura múltiples

trabajos que abordan la explicabilidad (*explainability*) del modelo. No existe una noción de explicabilidad unánime en la comunidad. Por ejemplo, en Vilone and Longo (2021) se recopilan más de 30 nociones de explicabilidad en el contexto de la inteligencia artificial del tipo: *"El grado de confianza de un algoritmo de aprendizaje para comportarse sensatamente en general", "un algoritmo de aprendizaje para transferir nuevos conocimientos a los usuarios finales."*, etc. Junto a estas nociones, los autores recopilan más de 100 artículos en donde se aplica alguna de las nociones. En realidad, la noción más adecuada de explicabilidad dependerá de varios factores, como el escenario concreto, si se trata de un usuario final o desarrollador, si el sistema ofrece una explicación en formato textual, visual o en términos de pesado de variables, el tipo de problema que se quiere resolver, etc. Centrándose en el área de PLN, Zini and Awad (2022) revisan las técnicas de explicabilidad interna de los modelos de lenguaje desde la perspectiva del desarrollador. Por ejemplo, se intentan interpretar las *embeddings* de palabras, las estructuras recursivas de redes neuronales, o los mecanismos de atención. Sin embargo, en este proyecto nos centraremos en la explicabilidad desde la perspectiva del usuario final, es decir, no tanto en entender los procesos y los errores del modelo, sino más bien los fundamentos de las decisiones tomadas por parte del modelo con vistas al usuario final. Esto se ha denominado *outcome explanation problem* (Guidotti et al., 2018). Dentro de esto, nos interesan las aproximaciones de auto-explicación local (*local outcome self-explaining*), consistentes en que el propio modelo genera una explicación para el usuario en cada una de las respuestas (Arya et al., 2019; Danilevsky et al., 2020). Dentro de la categoría de *local outcome self-explaining*, podemos distinguir tres conjuntos de aproximaciones dependiendo del formato de salida según la categorización dada en las secciones anteriores, esto es, organizacional y anotación, lenguaje formal y lenguaje natural.

En primer lugar, en cuanto tareas organizativas y de anotación, la explicabilidad se analiza mediante visualización de componentes relevantes (*salience*). Esta aproximación consiste en mostrar rasgos u otros componentes interpretables del modelo que ofrezcan información sobre las decisiones tomadas por el sistema. Algunos de estos componentes pueden ser palabras o tokens que han tenido más peso en la decisión (Godin et al., 2018), alineaciones de términos de entrada y salida (Bahdanau et al., 2015; Ferrando et al., 2023), marcar secuencias o palabras en los textos originales, (Mullenbach et al., 2018; Carton et al., 2018; Voskarides et al., 2015; Sydorova et al., 2019), o detectar muestras de entrenamiento que han sido más determinantes en la decisión (Abujabal et al., 2017; Croce et al., 2019). En estos casos, la explicación puede evaluarse sobre referencias humanas aplicando métricas de clasificación o de anotación de secuencias. Por ejemplo, en Carton et al. (2018) se aplica precisión y cobertura (métricas de clasificación) para evaluar explicaciones en identificación de ataques personales, en donde la explicación consiste en la identificación de las palabras que justifican dicha clasificación.

El segundo conjunto incluiría trabajos en los que se presenta una explicación en formato estructurado (generación de lenguaje formal) creada por el propio sistema. Por ejemplo, reglas o plantillas empleadas en la decisión (Abujabal et al., 2017), elementos de un grafo de conocimiento (Pezeshkpour et al., 2019), o formas lógicas (Liang et al., 2016). En estos casos, la evaluación en la literatura es cualitativa en la mayoría de los casos. Esto se debe a que, por un lado, existen múltiples explicaciones igualmente válidas, y por otro lado, las métricas de evaluación estándar empleadas en clasificación no se pueden aplicar sobre este tipo de estructuras.

En un tercer conjunto, podemos incluir explicaciones en formato de texto generadas por el modelo (generación de lenguaje natural). Por ejemplo, ChatGPT ofrece justificaciones ante las preguntas de conocimiento general de los usuarios. El buscador *Perplexity*[4] ofrece justificaciones en formato texto de los documentos recopilados ante una consulta. En el contexto académico, también hay múltiples trabajos centrados en problemas más específicos, como son explicaciones en resolución de problemas matemáticos (Ling et al., 2017), o cuestiones de sentido común (Rajani et al., 2019). En los casos en los que se ha evaluado cuantitativamente estas aproximaciones ha sido mediante métricas de texto libre como BLEU o ROUGE basadas en solapamiento de palabras con un texto de referencia. Sin embargo, este tipo de evaluaciones resulta aun más complejo que en el caso de otras tareas de generación de texto como traducción o resumen automático, dado que el conjunto de posibles explicaciones correctas es aún mayor.

[4]https://www.perplexity.ai

Dado que la transparencia de la explicabilidad es muy dependiente de la percepción del usuario, lo más fiable, aunque también más costoso, es emplear juicios humanos para evaluar la explicabilidad. Dada la subjetividad del problema, en algunos trabajos se han empleado hasta 25 jueces (Sydorova et al., 2019). También es necesario definir con cuidado las indicaciones proporcionadas a los jueces, dado que existen múltiples interpretaciones de la noción de explicabilidad (Vilone and Longo, 2021), que además puede adquirir matices propios en cada tipo de problema.

Evaluación de la explicabilidad en ODESIA

La explicabilidad en el año 2 del proyecto se evalúa dentro del Ámbito de Experiencia de usuario mediante cuestionarios sobre la justificación de respuestas que proporcionan los asistentes virtuales.

3.2. Competencias cognitivas

En torno al 2017, adquieren popularidad los modelos pre-entrados con grandes colecciones de texto y entrenados de nuevo sobre un conjunto de ejemplos de un problema en específico (*fine tuning*, *few shot learning*). Éstos consiguen ser muy efectivos en múltiples tareas. Poco después, surgirían modelos de lenguaje capaces de resolver problemas complejos sin muestras de entrenamiento, a partir de una especificación detallada del problema en la entrada (*prompting*). A partir de ese momento, la comunidad comienza a preguntarse no solo por la efectividad de los sistemas en problemas específicos, sino por sus competencias a nivel interno. Algunas de estas competencias son la capacidad de generalización, abstracción, razonamiento lógico, argumentación, productividad lingüística, composicionalidad, sistematicidad, etc. El análisis de estas competencias tiene efectos prácticos en el desarrollo de sistemas. Por ejemplo, se ha argumentado que la ausencia de *generalización composicional*, es decir, la construcción del significado de frases en función del significado de sus constituyentes y reglas de combinación, es el motivo por el que los sistemas necesitan tanta información de entrenamiento para generalizar correctamente (Lake et al., 2016). En la siguiente sección describimos los tres tipos de competencia que estructuran nuestro análisis. En las siguientes secciones se describe en detalle cada una de ellas y la bibliografía relacionada.

3.2.1. Tipos de competencia

En general, existe bastante ambigüedad en todos los términos que se refieren a competencias cognitivas, y son tratados de manera diferente por cada autor. Por ejemplo, Ziyu et al. (2023) categorizan las competencias de los modelos de lenguaje en: conocimiento lingüístico (semántico, gramatical, pragmático, etc) y del mundo, razonamiento causal, inductivo, abductivo, deductivo, basado en analogía, matemático, etc. En el contexto de las ciencias cognitivas, estos conceptos son difusos y dependen de la interpretación del autor. Por ejemplo, no resulta sencillo establecer límites entre *conocimiento semántico* y *conocimiento del mundo*, dado que la semántica se refiere a la conexión entre símbolos y referentes. También resulta difusa la frontera entre tipos de razonamiento como el matemático y el deductivo.

En este documento establecemos una categorización de competencias cognitivas de los sistemas basados en modelos de lenguaje sobre una base formal. Con este propósito, estructuramos las competencias en tres categorías que formalizamos mediante la noción de *significado*. Existen diferentes teorías sobre el significado en lingüística (Speaks, 2021), como son la teoría referencial, la teoría del uso del lenguaje, la ideacional, etc. Aunque la noción de significado es vaga, lo que sí podemos establecer de partida es que una expresión o texto no especifica completamente su significado, dado que siempre podríamos ampliar o matizar dicho texto. Es decir, elementos extra textuales como la experiencia previa del emisor o el contexto determinan el significado. Por tanto, podemos afirmar que tras una expresión existe un conjunto posible de significados a nivel pragmático (Stalnaker, 1999), independientemente de bajo qué teoría definamos la noción de significado. Podemos, entonces, partir de la idea de dos espacios: el de cadenas de palabras y el de significados. Las expresiones más vagas cubrirán un conjunto más amplio en el espacio de significados, mientras que las expresiones más específicas, largas o informativas tendrán asociado un conjunto más reducido[5].

[5]Es importante destacar que los conjuntos de significados no son discretos. Es decir, si un mensaje y todo su contexto

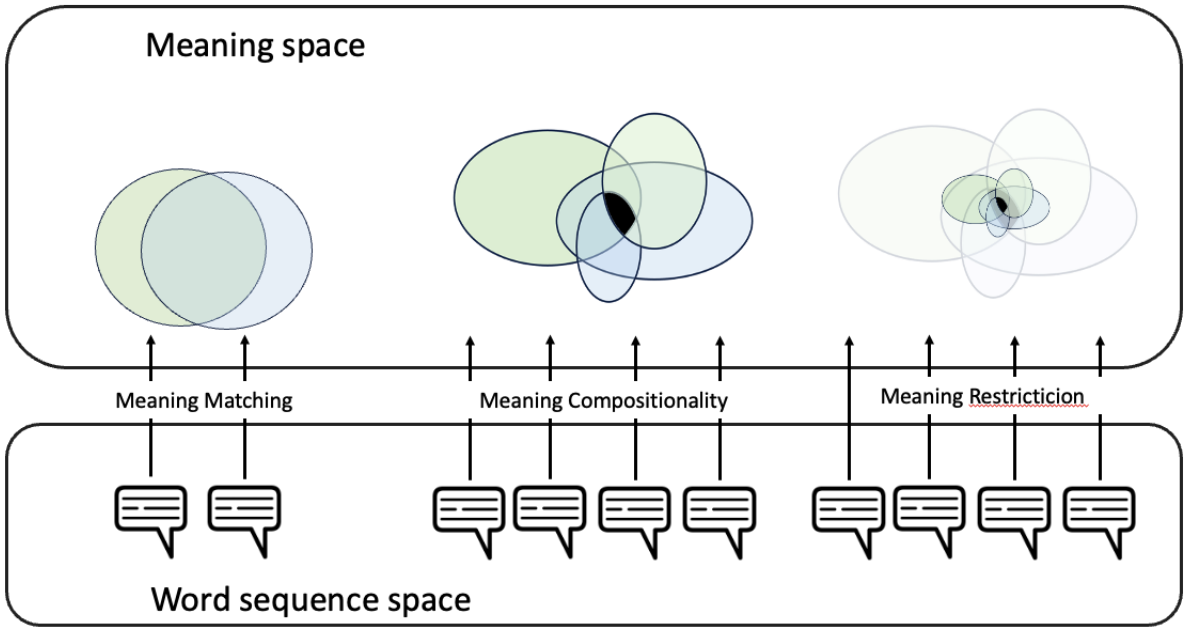

Figura 6: Categorización de competencias lingüísticas en base al tratamiento de significados asociados a cadenas de palabras.

La Figura 6 ilustra esta idea. En la parte inferior se encuentra el espacio de cadenas de palabras. Cada cadena tiene una correspondencia con un conjunto de significados posibles en el espacio superior. Por ejemplo, tras la expresión *"Calentaré una sopa"* están todos los significados que puedan existir detrás de dicho mensaje. Lo mismo ocurriría con la expresión *"Tengo hambre"*. La intersección entre ambos conjuntos representaría los significados que pueden encontrarse simultáneamente tras las expresiones *"Calentaré una sopa"* y *"Tengo hambre"*. Es decir, en cierto contexto situacional o pragmático adecuado, el emisor podría lanzar cualquiera de los dos mensajes si lo que desea es comunicar que calentará una sopa porque tiene hambre, aunque ninguno transmita de manera completa la información. Imaginemos que alguien llega a las tres de la tarde a casa y dice "Tengo hambre" o "Calentaré una sopa" mientras saca una olla de la nevera. En ese contexto ambos enunciados cobran sentido. Para estandarizar el lenguaje, en lo que sigue en este documento hablaremos de *significado de un texto* para referirnos al conjunto de significados posibles a nivel pragmático en dicho espacio.

El significado puede atribuirse tanto a cualquier unidad lingüística (morfemas, palabras, expresiones, textos), como a categorías. Por ejemplo, las categorías definidas en un problema de clasificación (etiquetas de clase) tienen su significado. La categoría *biología*, es un concepto abstracto que puede especificarse en el significado de un documento de biología concreto. La intersección entre los conjuntos de significados asociados a las categorías *biología* y *física* incluirá el significado de cualquier documento que incluya ambas temáticas. La unión representará cualquier documento que incluya al menos una de las dos.

Para analizar las competencias de los modelos de lenguaje, proponemos distinguir entre *ajuste*, *composición* y *restricción* de significados en función de la implicaciones del problema en términos de conjuntos de significados. Es importante tener en cuenta que no estamos haciendo ninguna disquisición sobre los procesos cognitivos empleados durante la realización de esta tarea, sino el hecho de que exista o no una

tiene asociado un significado, el mensaje (sin contexto) se corresponderá con una distribución de probabilidad en el espacio de significados. Es decir, cada elemento en el espacio de significados tendrá una probabilidad (función de densidad) de encontrarse tras la expresión. Esto se traduce en un conjunto difuso tal que los grados de pertenencia de los significados al mensaje suman uno. La unión, intersección de conjuntos no tiene por qué cumplir esta propiedad. Podemos interpretar los conjuntos discretos como umbrales de probabilidad en la distribución.

correspondencia entre conjuntos de significados.

- Definimos la competencia de **ajuste de significados** como *la realización de tareas que se traducen en la identificación de significados similares asociados a expresiones distintas.* Por ejemplo, asociar las expresiones *"Cómo hacer si uno se pasa con el azúcar cuando hace una tarta de manzana"* y *"Mitigar el exceso de azúcar en la preparación de una tarta de manzana"* supone asignar dos secuencias de palabras a un mismo (o similar) conjunto de significados. Una pregunta de desarrollo en un examen del tipo *"Revolución industrial en Inglaterra"* cuyo contenido se encuentre en un capítulo en un libro de texto, se correspondería con un problema de similitud de conjuntos de significados asociados a la pregunta y a la respuesta. Con el fin de usar una nomenclatura más estándar, nos referiremos al ajuste de significados como *variación lingüística*, que comúnmente se define como la semejanza de significados *asociados* a distintas expresiones o referentes.

- Definimos la competencia de **composicionalidad de significados** como *la realización de tareas que se traducen en aplicación de operadores entre significados asociados a la entrada y a los datos de entrenamiento*. Por ejemplo, detrás de la expresión *Mitigar el exceso de azúcar* puede haber un significado relativo a cualquier receta, o incluso al exceso de azúcar en sangre. Detrás de *"preparación de una tarta de manzana"* puede haber cualquier significado relacionado como el título de una receta o un punto en las tareas del día. La intersección entre los dos significados contendrá el significado asociados a la expresión *"Mitigar el exceso de azúcar en la preparación de una tarta de manzana"*. Pero la composicionalidad de significados también incluye generalización o unión de conjuntos. Por ejemplo, supongamos que el sistema dispone de información sobre *preparación de una tarta de chocolate* y *preparación de una tarta de fresa*. En este caso, podría generalizar el concepto de *preparación de una tarta* para luego aplicar su intersección con los significados asociados a *Mitigar el exceso de azúcar* (ver gráfico del centro en la Figura 6). Por ejemplo, una pregunta de examen como *"Clases pobres en la Inglaterra de la revolución industrial"* puede corresponderse con la generalización (unión de significados asociados a documentos sobre *clases obreras en Inglaterra* y *agricultores en Inglaterra* y su intersección con *revolución industrial en Inglaterra*. La composición de significados está estrechamente relacionada con el principio de composicionalidad en semántica (Szabó, 2022) y con la noción de generalización composicional. Se ajusta a la situación en la que la información que debe devolver el sistema se encuentra distribuida en varios fragmentos de texto.

- Definimos la competencia **restricción de significados** como *la realización de tareas que se traducen en operadores entre conjuntos restringidos de significado*, entendiendo como conjuntos restringidos a un subconjunto de los posibles significados de los mensajes, es decir, conjuntos en el espacio de significados resultantes de un modelado del problema, o lo que es lo mismo, una interpretación del mundo o abstracción (gráfico de la derecha de la Figura 6). La noción de conjunto restringido de significados está estrechamente relacionada con el razonamiento y la inferencia. Esta competencia se ajusta a la situación en la que la información que debe devolver el sistema no se encuentra en ningún lugar ni distribuida en textos, sino que es necesario restringir los significados estableciendo relaciones de implicación dependientes del problema.

- Por último, en un extremo podríamos añadir la competencia que se corresponde con el **razonamiento matemático** (en la que no profundizaremos), en la que no se manejan conjuntos, sino puntos infinitesimales en el espacio de significados. Esto tiene diversas consecuencias. Es la base de las ciencias exactas y requiere la producción conjuntos infinitos de símbolos (números) para representar infinitos significados.

En resumen, la variación lingüística supone estimar similitudes entre conjuntos de significados asociados a los textos, la composición supone además operar con estos conjuntos, y la restricción de significados implica la definición de conjuntos (abstracción).

Partiendo de este contexto, podemos concluir que los niveles de competencia no están asociados al problema en sí, sino a la información de la que se dispone en el corpus de entrenamiento. Por ejemplo,

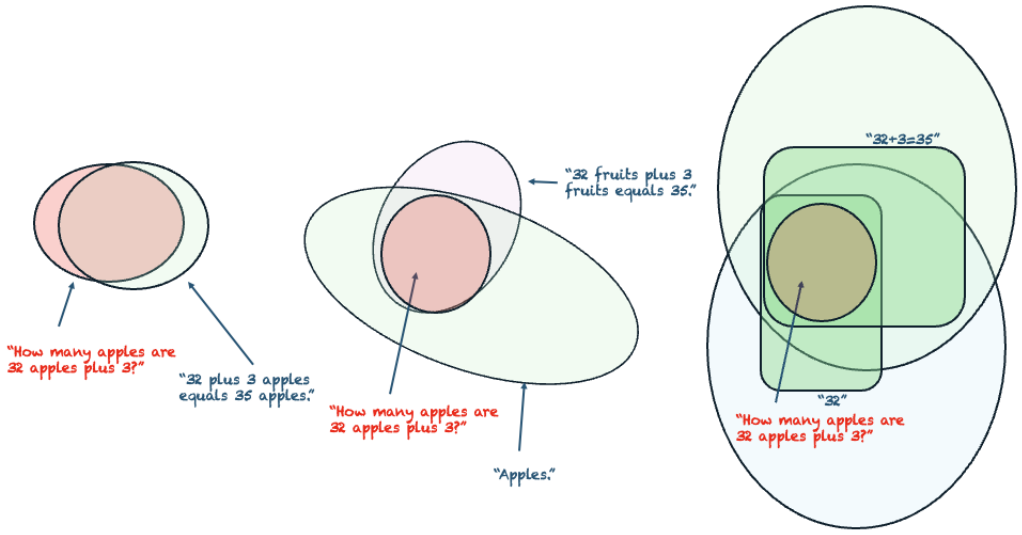

Figura 7: Ejemplo de competencias lingüísticas sobre los espacios de cadenas de palabras y significados.

supongamos que se le pide al sistema resolver la pregunta *"¿Cuántas manzanas son 32 manzanas más 3?* (Ver figura 7). Si el corpus contiene una afirmación del tipo *"32 más 3 manzanas son 35 manzanas"* estamos ante un problema de ajuste de significados o variación lingüística. Si contiene una afirmación del tipo *"32 más 3 naranjas son 35 naranjas"* estamos ante un problema de composición (vía generalización). Si el corpus contuviera conceptos abstractos como números *(32, 3, 35)* y relaciones del tipo *32+3=35*, entonces la cuestión no se podría solucionar sin un proceso de abstracción o modelado del problema. Es decir, restringir el conjunto de significados de *33, 3, 35*, y *32+3=35* al contexto de la acumulación de manzanas.

Por definición, un modelo de lenguaje no modela el espacio de significados, por lo que no se da en realidad ninguna operación de unión o intersección de conjuntos, ni tampoco ningún proceso de abstracción. Sin embargo, podemos evaluar la capacidad de realizar tareas en las que se establecen implícitamente estos tipos de operaciones mediante el modelado probabilístico del lenguaje.

3.2.2. Ajuste de significados: Variación lingüística

Desde los primeros años, la versatilidad del lenguaje ha sido la principal barrera a la que se ha enfrentado el procesamiento de lenguaje natural. A pesar de la disponibilidad de grandes colecciones de documentos en formato digital, basta combinar unas pocas palabras entrecomilladas en un buscador para que éste no encuentre ni una ocurrencia de la mismas en el mismo orden. Esto se debe a que un mismo significado o significados similares pueden representarse de innumerables formas con palabras. El proceso mediante el cual se supera esta barrera es lo que comúnmente se denomina *generalización del lenguaje.* Tradicionalmente, se distingue entre tipos de generalización dependiendo de los niveles del lenguaje (léxico, sintáctico, semántico, discursivo, etc.). En el contexto de los sistemas de procesamiento de lenguaje, hemos definido la competencia de *ajuste de significados* como *la realización de tareas que se traducen en proximidad de significados de fragmentos de texto, ya sean de la entrada o de la colección de entrenamiento*. Es decir, el significado del texto de entrada o fragmentos del corpus de entrenamiento es similar, variando únicamente el texto empleado para describirlo. Esta noción encaja perfectamente con lo que se entiende como *variación lingüística*, en donde diferentes formas expresan un mismo significado. El caso paradigmático de variación lingüística en el contexto de sistemas inteligentes es la recuperación de información. Por ejemplo, un buscador web con competencia de variación lingüística debe ser capaz de asociar la consulta del usuario con algún fragmento de algún documento de la colección. Pero también,

un clasificador de textos devuelve una respuesta ante un nuevo texto por su similitud con alguna muestra del conjunto de entrenamiento.

En la lingüística computacional clásica previa a los 90, la variabilidad lingüística se aborda desde las gramáticas y expresiones regulares, es decir, mediante sistemas basados en reglas. A lo largo de la década de los 90 se desarrollaron ontologías como WordNet (Miller, 1994) para el inglés, que luego se extendería a otras lenguas. Estas ontologías léxicas consisten en una jerarquía de sentidos a los que se les asocia una terminología (sinónimos). La mayoría de los términos del vocabulario se ajustaban a varios sentidos dependiendo de su contexto, por lo que se desarrolló a partir de este recurso una larga línea de trabajos en desambiguación semántica (Navigli, 2009). En paralelo, surgen en esta época las representaciones vectoriales de textos basadas en frecuencia de palabras y la variabilidad lingüística se modela mediante medidas de similitud en el espacio vectorial. A lo largo de la década de los 2000, se aborda el problema de la variabilidad lingüística a nivel de frase. Es decir, se plantea como tarea a resolver el reconocer pares de frases o párrafos que comparten significado. Esto dio lugar, además de múltiples trabajos, a la participación de grupos de investigación en las campañas de evaluación *Semantic Textual Similarity* en las conferencias SEMEVAL (Chandrasekaran and Mago, 2021; Cer et al., 2017a). También destacan en este período métodos de reducción de dimensionalidad como *Latent Semantic Indexing*, así como métodos de *Topic Modeling*.

Finalmente, en la década de los 2010 toman protagonismo los modelos neuronales en donde se fusiona la representación vectorial (estados de activación de la red ante una entrada) con la potencia predictiva de los modelos de lenguaje (pre-entrenamiento de la red basado en predicción de secuencias de palabras). Los primeros modelos como Word2Vec o Glove abordaban la variabilidad léxica mediante representaciones estáticas de palabras. Esto da paso a los modelos de lenguaje neuronales actuales, que abordan la variabilidad lingüística a nivel de frase o documento (Brown et al., 2020; Rae et al., 2021).

Como hemos apuntado anteriormente, la recuperación de información puede interpretarse como un problema de ajuste de significados donde es necesario conectar el conjunto de significados asociado a la consulta con el conjunto de significados de cada documento de la colección. La aplicación clásica de los modelos de lenguaje son las técnicas de *query likelihood language models* en donde se infiere un modelo de lenguaje por cada documento para luego medir el *likelihood* de la consulta en cada uno de ellos (Zhai, 2008). Sin embargo, en los años recientes se ha incorporado el uso de modelos de lenguaje pre-entrenados sobre toda la colección. En Zhu et al. (2023) los autores distinguen entre el uso de estos modelos para la reescritura de consultas, la búsqueda, el *re-rankeo* de documentos o el proceso de lectura. En cualquier caso, la aplicación de los modelos de lenguaje en estos contextos se centra en la fase de representación. Es decir, el uso de métricas de similitud entre la representación vectorial de la consulta y la representación de los documentos de la colección. Estas representaciones vienen dadas por el estado de activación de la red entrenada al introducir los documentos o la consulta. La literatura muestra en general una alta competencia en cuanto a variación lingüística en este contexto (Zhu et al., 2023).

Existen pocos estudios en donde se haya evaluado la competencia de variación lingüística o ajuste de significados de manera aislada en los modelos de lenguaje neuronales en un contexto generativo. En el corpus TriviQA de búsqueda de respuestas, se categorizan casos de test en donde se pone a prueba aspectos de la variabilidad lingüística como *variabilidad léxica*, *variabilidad sintáctica*, frente a otras categorías de casos de test que van más allá de la variabilidad lingüística a las que los autores denominan *conocimiento del mundo* o *frases múltiples* (Joshi et al., 2017). A veces un modelo neuronal puede sobre-generalizar la variación lingüística identificando correspondencias semánticas falsas entre textos similares. Para evaluar este problema el conjunto de datos de preguntas de sentido común WINOGRANDE se generan casos de test con textos similares a nivel léxico y sintáctico (Sakaguchi et al., 2021).

En general, evaluar de manera aislada la generalización por variabilidad lingüística requiere definir casos de test aportando al sistema muestras de entrenamiento o fuentes con la solución expresada de manera distinta. La complejidad de evaluar la variabilidad lingüística está en asegurar que el sistema no tiene acceso a fragmentos expresados en los mismos términos que la entrada al sistema.

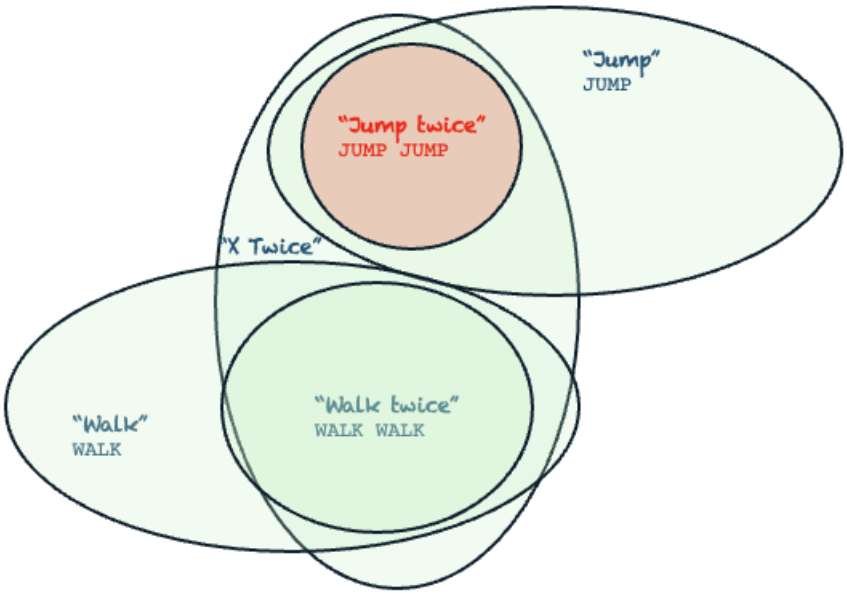

Figura 8: Ejemplo de competencias de composición de significados.

3.2.3. Composición de significados

En la sección 3.2.1 definimos la competencia de composición de significados como *la realización de tareas que se traducen en proximidad y operadores de conjunto entre significados asociados a fragmentos, bien de la entrada, o bien del corpus de entrenamiento*. La idea de composición de significados está íntimamente relacionada con algunas nociones en lingüística. Un de ellas es la noción de *generalización composicional*, que ha sido definida como *"la capacidad de extender el aprendizaje a través de la combinación de elementos individuales en una estructura más compleja"* (Fodor and Lepore, 2002). En el contexto de los modelos de lenguaje, Shaw et al. lo definieron como *"la habilidad para generalizar a combinaciones novedosas de los elementos observados durante el entrenamiento ´´* (Shaw et al., 2021). Gu et al. (2020) definen la generalización composicional en el contexto de generación de lenguaje formal como el procesamiento de texto cuando la expresión formal correspondiente no se encuentra explícitamente en los datos de entrenamiento.

Otra noción relacionada con la composición de significados es el principio de composicionalidad en lingüística, es decir la idea de que el significado de expresiones completas viene dado por el significado de sus constituyentes y la forma en que se relacionan sintácticamente entre sí (Baroni, 2019). Diversos autores han intentado desgranar la generalización composicional en operaciones composicionales como *sistematicidad*, *productividad*, *sustitución*, *localismo* y *sobregeneralización* (Hupkes et al., 2021). Dankers et al. (2022) distingue entre *sistematicidad*, *sustitución* y *composición global*. An et al. (2023) distingue entre *sustitución de primitivas*, *alternancia de estructuras de primitivas*, *combinación de sintagmas*, etc. En el corpus de preguntas SQUADRUN se categorizan casos de test en fenómenos como *negación*, *antonimia*, *sustitución de entidades*, *exclusión mutua* o *condiciones imposibles* (Rajpurkar et al., 2018). Por otro lado, la generalización composicional representa toda una línea de investigación en lingüística, dando lugar a múltiples teorías y falta de consenso en la comunidad.

Un ejemplo de competencia composicional recurrente en varios artículos de la bibliografía (Bastings et al., 2018) es interpretar *"jump twice"* cuando en el corpus de entrenamiento solo se encuentra *"walk"* *"jump"* y *"walk twice"* (Figura 8). En este caso, la resolución del problema se traduce en considerar la generalización (unión) de todos los conjuntos de significados asociados a un verbo X seguido de *"twice"*

(i.e. *"X twice"*). De alguna manera, si el modelo de lenguaje tiene competencia composicional, entonces aun no existiendo la secuencia *"jump twice"*, puede generalizar la secuencias *"X twice"* e intersectarla con *"jump"*.

En general, el grado de competencia en composicionalidad de los modelos de lenguaje es aun un tema de discusión en la comunidad. Por ejemplo, diversos trabajos reportan carencias composicionales en los modelos estudiados (Lakretz et al., 2019; Gulordava et al., 2018; Lake and Baroni, 2017). Por lo contrario, otros autores afirman que ciertas relaciones composicionales pueden ser aprendidas durante el proceso de entrenamiento (Valvoda et al., 2022; Bastings et al., 2018).

La composicionalidad de significados se ha evaluado de manera indirecta mediante conjuntos de datos para problemas que requieren dicha competencia. Algunas tareas que requieren composicionalidad de significados son la generación de código de programación (Chen et al., 2021; Austin et al., 2021a), lectura y comprensión (Rajpurkar et al., 2016a; Choi et al., 2018; Trischler et al., 2017), respuestas en formato largo (Long-Form Question Answering) (Qin et al., 2023), o inferencia textual (Goodwin et al., 2020)). En estos marcos de evaluación se intenta asegurar que la solución del problema requiera operaciones de composición.

En lo que respecta a tareas de generación de código, se puede asegurar que el sistema debe hacer uso de competencias de composicionalidad siempre que la solución al problema propuesto no se encuentre en los datos de pre-entrenamiento. Por ejemplo, la traducción de un bucle de uno a 115 a python se obtiene por generalización de bucles de 1 a n y particularización a 115. Por ejemplo, el corpus HumanEval (Chen et al., 2021) incluye especificaciones del tipo *"dada una lista de enteros, devolver la suma de elementos que se encuentran en posiciones pares."* Es muy posible que estos códigos no existan explícitamente en el corpus de pre-entrenamiento, aunque sí códigos semejantes que es posible generalizar. La evaluación se lleva a cabo con métricas de generación de código descritas en la sección 3.1.1.

En cuanto a los conjuntos de test de lectura y comprensión, el sistema debe identificar la respuesta correcta dentro de un texto. Algunas de las competencias que se plantean en estos conjuntos tienen que ver más con la variación lingüística que con la composicionalidad, como la sinonimia, variación léxica, o variación sintáctica, pero también incluyen competencias de composicionalidad como la resolución de anáfora o fusión de múltiples frases (Rajpurkar et al., 2016a). El conjunto de test NewsQA también incluye respuestas que requieren el solapamiento de conceptos o la síntesis de información a partir de diferentes pasajes (Trischler et al., 2017). La calidad de las respuestas se evalúa mediante métricas de anotación de secuencias, tasa de aciertos o medida F entre conjuntos de términos.

Algunas tareas de búsqueda de respuesta también requieren composición. Por ejemplo, en los casos en los que la tarea requiere tanto recuperación de información como síntesis (Long-Form Question Answering) (Krishna et al., 2021; Qin et al., 2023). La evaluación se lleva a cabo mediante métricas de generación de texto tipo ROUGE descritas en la sección 3.1.1.

Sin embargo, en los trabajos anteriores, no se estudia de manera específica y aislada la competencia composicional de los modelos. Para ello se han presentado diferentes metodologías. En la mayoría de los estudios se evalúan los modelos sobre un conjunto de datos sintéticos en donde se diseñan muestras de entrenamiento y test específicos para evaluar relaciones de composicionalidad específicas como las descritas al comienzo de esta sección (Hupkes et al., 2021; An et al., 2023). Uno de los conjuntos de datos de evaluación más empleados es SCAN (Lake and Baroni, 2017), consistente en expresiones en lenguaje natural (*"Jump twice"*) asociadas a secuencias de órdenes (`JUMP JUMP`), probablemente inspiradas en el escenario de un interfaz de usuario (*pasa pantalla*, *cierra el fichero*) o en robótica (*camina, gira*). En NACS se añade un corpus complementario a SCAN en donde se requiere un mapeo a la inversa desde las secuencias de órdenes a las expresiones en lenguaje natural (Bastings et al., 2018).

Más allá de la generación de secuencias de instrucciones, en ciertos conjuntos de datos como CFQ (Keysers et al., 2020) Y COGS (Kim and Linzen, 2020), la salida del sistema se expresa en lenguaje formal. En CFQ, además de preguntas y respuestas en lenguaje natural, se proporciona para cada pregunta, una consulta SPARQL correspondiente contra la base de conocimientos Freebase (Bollacker et al., 2008). En COGS las salidas consisten en estructuras semánticas basadas en lambda-cálculo, es decir, un sistema de cláusulas y proposiciones lógicas en un dominio restringido (Kim and Linzen, 2020). Por ejemplo, el

modelo es entrenado para traducir expresiones como *"A cat smiled"* en la expresión formal cat(x_1) AND smile.agent(x_2,x_1). En estos conjuntos de test los sistemas son evaluados en base a la tasa de aciertos.

En ciertos conjuntos de datos como CFQ (Keysers et al., 2020), COGS (Kim and Linzen, 2020), el data set del modelo RulerTaker (Clark et al., 2020) o de CriPT (Betz et al., 2021), las muestras son generadas mediante gramáticas. Es decir, las expresiones de entrenamiento y test no se escriben a mano ni se recolectan, sino que se generan a partir de un sistema de reglas. En otras palabras, se explotan las gramáticas desarrolladas en el paradigma clásico de la lingüística computacional para la construcción de conjuntos de datos de entrenamiento. Saxton et al. (2019) enumera las diversas ventajas de producir datos de entrenamiento mediante reglas, como son la escalabilidad, la eliminación de errores humanos o el control del dominio. En el caso de COGS, los autores generan reglas para: *combinación de primitivas y roles gramaticales*, *combinación de sintagmas modificados y roles gramaticales*, *recursividad*, *alternancia de argumentos verbales* y *clases de verbos.*

Sin embargo, algunos autores han reportado la importancia de considerar conjuntos de datos no sintéticos al evaluar la competencia composicional, es decir, muestras de entrenamiento y test representativas de casos de uso reales (Shaw et al., 2021; Keysers et al., 2020). Para capturar la composicionalidad, estos autores proponen la división de muestras de aprendizaje y test de forma que el muestreo sea representativo pero asegurando la necesidad de composicionalidad.

En cuanto a metodologías de evaluación de competencias composicionales, a la vista de estos trabajos y dado un escenario o problema específico, podemos considerar como recomendables las siguientes metodologías dependiendo del tipo de problema:

- **Generación de lenguaje formal en dominio controlado.** En este tipo de escenarios resulta apropiado generar un conjunto de datos representativo mediante el uso de reglas generativas, capturando la efectividad del sistema en relación a diferentes tipos de generalización composicional. El modelo puede evaluarse en función de la validez de las estructuras formales generadas o también en función de la cantidad de información textual que el sistema necesita para generar dichas estructuras.

- **Tareas organizacionales o generación de texto libre en dominio abierto.** Por ejemplo, la generación de respuestas con texto libre, ejercicios de lectura y comprensión o resolución de cuestionarios que requieran la síntesis de información distribuida en distintos fragmentos. En este caso es más apropiado el uso de muestras reales que capturen las características del problema. Para asegurar la competencia composicional se puede realizar una división de muestras de aprendizaje y test de forma que se asegure la composicionalidad. En cuanto a métricas de evaluación, se pueden aplicar métricas de clasificación o ranking o bien medidas de similitud entre salidas y textos de referencia.

Al igual que en la competencia de ajuste de significados, en ambos casos es importante asegurar que el modelo no tiene acceso a respuestas que no requieran el tipo de generalización que se está evaluando. Concretamente, las muestras de entrenamiento y las fuentes no deben contener una solución a la que se pueda acceder solo por variación lingüística, sino que debe ser necesario componer significados. Es lo que denominaremos *contaminación de los datos de test.*

3.2.4. Restricción de significados: Razonamiento e inferencia.

Hemos definido la restricción de significados como *la resolución de problemas que requieren relaciones de implicación entre significados restringidos al problema*. Esto quiere decir que la salida del sistema no puede obtenerse por intersección o generalización de significados de fragmentos de textos, sino que es necesario restringir dichos significados en función del problema. Tomemos por ejemplo, el caso de inferencia textual del corpus e-Care (Du et al., 2022): *"Tom sostiene un bloque de cobre con la mano y lo calienta en el fuego"* implica *"Siente ardor en sus dedos inmediatamente"* (Figura 9). Es posible que esta implicación aparezca textualmente expresada en algún fragmento sobre el que ha sido entrenado el sistema. También es posible que pueda derivarse a partir de composición de significados, dependiendo de lo significados que se deriven del corpus de entrenamiento, los modelos de lenguaje y la semántica distribucional. Pero en principio, para identificar la relación de implicación entre conjuntos de

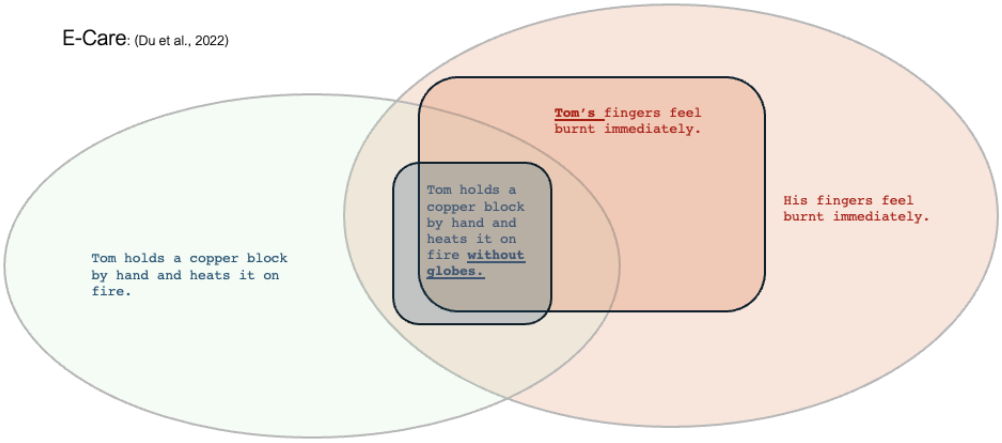

Figura 9: Ejemplo de restricción de significados.

significados de ambas expresiones es necesario particularizarlas. Por ejemplo, dentro del conjunto posible de significados de la primera frase, tendríamos que asumir que Tom no lleva guantes. Dentro del conjunto de significados asociados a la segunda expresión, es necesario particularizarlo en base a la anáfora. Es decir, el sujeto se refiere a Tom. En otras palabras, para extraer relaciones causales o de implicación es necesario realizar asunciones sobre el significado de las expresiones, es decir, interpretar el problema.

La restricción de significados tiene una estrecha relación con la noción de *razonamiento*. Se entiende por razonamiento a la facultad que permite resolver problemas, extraer conclusiones y aprender de manera consciente de los hechos, estableciendo conexiones causales y lógicas necesarias entre ellos. Según la enciclopedia británica, se define como *"El proceso de pensar en algo de manera lógica para formar una conclusión o juicio."*. Estas relaciones causales requieren un modelado del problema y por tanto, asunciones que restringen el significado de las expresiones. También está relacionado con la noción de *semántica formal* en la que es necesario un modelado del mundo, en contraposición con *semántica distribucional* en donde se modela la coocurrencia de símbolos del lenguaje (Venhuizen et al., 2019).

Existen muchas categorías y categorizaciones distintas del proceso de razonamiento en ciencias cognitivas. Por ejemplo, razonamiento lógico, argumentativo, deductivo, abductivo, crítico, decomposicional, etc. Se suele distinguir entre razonamiento argumentativo y razonamiento lógico. En el segundo, las relaciones de implicación son definidas a priori por lo que se puede aplicar de manera mecánica las reglas lógicas para derivar la verdad o falsedad de proposiciones. En el caso del razonamiento argumentativo sin embargo, las relaciones de implicación están sujetas al contexto o significados posibles de las expresiones. En (Davis, 1990) de definen estas dos categorías como razonamiento lógico frente a sentido común, el cual requiere comprensión del mundo (Davis, 1990).

Se puede establecer una correspondencia entre operadores lógicos y relaciones entre conjuntos de significados, ya sean en el espacio de significados reales o en un espacio definido para dicho problema en cuestión. La Figura 10 muestra dicha relación. La proposición de verdad o falsedad puede interpretarse como la inclusión de un conjunto de significados dentro de un contexto. Por ejemplo, en el contexto de significados *con sentido* de Harry Potter, *"Las cartas vuelan"* pertenece al conjunto de significados dotados de verdad. Los operadores lógicos de unión, intersección o implicación también tienen una traducción directa a operadores de conjuntos en el espacio de significados. También las proposiciones con argumentos pueden traducirse. En el gráfico de la esquina inferior derecha en la figura, el conjunto más amplio cubre todos los posibles significados asociados a la instancia x, por ejemplo, Goofy dentro del espacio X de personajes de Walt Disney. Dentro de todos estos posibles significados, la proposición $A(x)$ podría representar por ejemplo, que el personaje x aparece en cierta película. El gráfico de la esquina superior derecha representa las relaciones proposicionales. Una relación entre x e y se encuentra dentro de

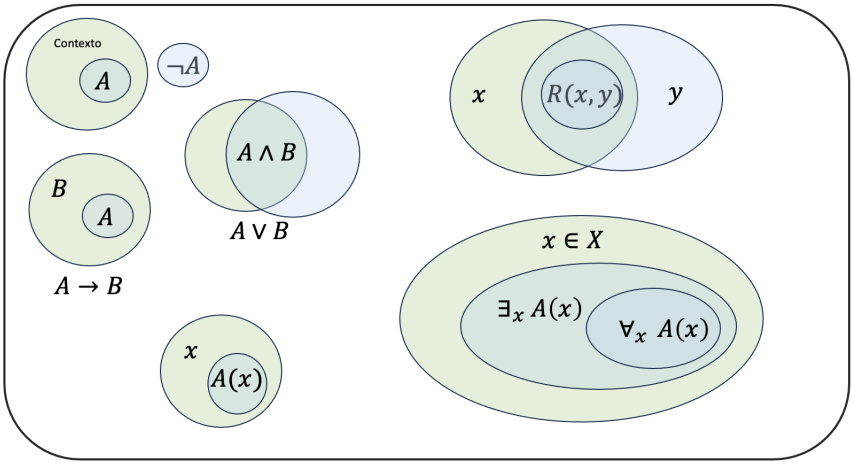

Figura 10: Operadores lógicos sobre conjuntos de significados.

los conjuntos posibles de significados asociados simultáneamente a x e y, y dentro de estos, los posibles significados en donde se establece una relación R entre ambos. Por último, también los operadores de cuantificación son interpretables en términos de conjuntos de significados, como muestra el gráfico inferior derecha de la figura.

En la literatura encontramos múltiples estudios que estudian la competencia de razonamiento en modelos de lenguajes. Siguiendo el análisis de Guo et al. (2023), estos tests o tareas pueden agruparse en tres conjuntos dependiendo de la información de entrada y de salida del sistema.

Categorización de relaciones lógicas. Esta tarea consiste en identificar relaciones lógicas del tipo implicación, contradicción o neutralidad entre pares de expresiones en lenguaje natural. Es decir, la entrada consiste en pares de frases y la salida una etiqueta de tipo de relación entre frases a nivel de inferencia. Esta tarea tiene ya algunos años de historia en los que se han desarrollado diferentes conjuntos de datos anotados. En los primeros años el problema se centró en la implicación textual en donde se distingue entre relaciones de implicación, contradicción y neutralidad data una hipótesis y una conclusión (Poliak, 2020). Dos de los conjuntos más populares son el Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) y su sucesor Multi-NLI (Williams et al., 2018), que incluyen en torno a medio millón de ejemplos. Algunos conjuntos más recientes ha centran el tipo de casos. Por ejemplo, el conjunto ConjNLI se centra en implicación textual entre frases conjuntivas (Saha et al., 2020). El conjunto HELP se centra en casos de *razonamiento monótono* en donde la implicación textual viene dada por sustitución de palabras (Yanaka et al., 2019).

Existen conjuntos de datos anotados de relaciones a tres frases, en concreto, relaciones de causalidad. Este es el caso del conjunto de datos CausalBank (Li et al., 2021b), así como en conjuntos de datos de razonamiento abductivo, en donde dada una hipótesis y una explicación, se ha de identificar la expresión que explica la conexión entre ambos (Bhagavatula et al., 2020).

Algunos conjuntos de test ofrecen una variedad más amplia de relaciones. Por ejemplo, en (Joshi et al., 2017) se define una taxonomía de relaciones distinguiendo entre inferencia lingüística (léxica, sintáctica, etc.), lógica (negación, conteo, causalidad, etc.) y conocimiento (taxonómico, conocimiento del mundo), siendo los dos últimos los correspondientes a razonamiento lógico y argumentativo respectivamente. La tarea consiste en identificar una de las 15 relaciones posibles. En estudio de los autores, el modelo RoBERTa llegó a obtener en torno a un 90 % de acierto en todas las categorías. En (Gu et al., 2023), sobre 100 objetos cotidianos, se pide a anotadores humanos que dibujen un gráfico que represente su modelo mental de partes, las relaciones espaciales, las conexiones entre sus partes y las dependencias funcionales

(si las hay). La entrada a al sistema consiste en una lista de objetos cotidianos, partes y 14 relaciones. La salida a generar consiste en una lista de tuplas (x, r, y) donde x a y mantengan una relación r. En este caso, la necesidad de un modelo mental y la variedad de relaciones hace que los mejores sistemas no superen un 50 % o 60 % de aciertos.

Otros conjuntos de test consisten en la detección de relaciones de implicación entre conjuntos de premisas e hipótesis. Por ejemplo, el conjunto de test FOLIO, incluye por cada entrada un conjunto de hipótesis expresadas en lenguaje natural y en lenguaje formal. El sistema debe decidir si las hipótesis son ciertas dadas las premisas (Han et al., 2022). El hecho de tener que validar varias hipótesis simultáneamente podría implicar ciertas competencias de razonamiento. Sin embargo, no puede descartarse tampoco la generalización sobre casos similares de inferencia formal.

A la hora de evaluar la identificación de este tipo de relaciones es muy importante evitar los sesgos en el muestreo. Por ejemplo, en (Bowman et al., 2015) se demostró que un sistema que observe únicamente la hipótesis puede llegar a obtener resultados medianamente aceptables. Saxton et al. demostraron que a pesar de los esfuerzos por reducir el problema, éste persiste en conjuntos de datos desarrollados tras su descubrimiento en 2018. En este sentido (Joshi et al., 2017) distingue entre datos sintéticos y muestras extraídas de un corpus de texto, siendo las segundas más representativas del lenguaje real y menos dadas a este tipo de problemas. Independientemente de este tipo de sesgos, estos test permiten comprobar si el modelo es capaz de inferir relaciones lógicas entre expresiones. Sin embargo, identificar estas relaciones no implica necesariamente competencias de razonamiento. En realidad, bastaría con cierta capacidad de generalización y suficiente información en los datos de pre-entrenamiento.

Resolución de cuestionarios. El segundo tipo de datos consiste en preguntas tipo test que requieren, al menos para un humano, competencias de razonamiento. La entrada consiste en una pregunta y un conjunto de respuestas candidatas. La salida es evaluada como un problema de clasificación. Encontramos en la literatura conjuntos de datos orientados a diferentes tipos de problemas como plausibilidad de sucesos (Wang et al., 2018c), razonamiento sobre acciones físicas (Bisk et al., 2020), razonamiento temporal sobre sucesos implícitos (Zhou et al., 2021), relaciones de causalidad (Du et al., 2022) o situaciones sociales (Sap et al., 2019).

La dificultad de definir estas preguntas radica en asegurar que la respuesta requiera que el sistema evaluado tenga competencias de razonamiento más allá de la composicionalidad de información de entrenamiento. Para ello, el conjunto de datos no debe estar *contaminado*, es decir, la solución no debe encuentre explícitamente en ningún lugar accesible en Internet. Además la respuesta debe requerir una cierta interpretación o modelado del mundo. Diferentes autores han aplicado diferentes criterios para asegurar la competencia de razonamiento y modelado del mundo. Por ejemplo, en C-Eval (Huang et al., 2023) se recopilan cuestionarios de examen de diferentes niveles de dificultad, de exámenes simulados o de exámenes de escuelas locales para evitar la contaminación. Sin embargo, no se hace un estudio sobre el grado de contaminación. En (Talmor et al., 2019) las preguntas se definen a partir de un grafo de conocimiento con relaciones entre conceptos para asegurar que la respuesta requiere el uso de sentido común. La dificultad de la tarea hace que el modelo BERT no supere el 56 % de acierto.

En el conjunto de datos OpenBookQA, se introducen preguntas que requieren combinar hechos de un libro de acceso libre (por ejemplo, los metales conducen electricidad) con un conocimiento común amplio (por ejemplo, una armadura está hecha de metal) extraído de otras fuentes (Mihaylov et al., 2018). Obtuvieron resultados bastante pobres mediante modelos de ese momento. En el conjunto de datos desarrollado en (Qin et al., 2021) la tarea consiste en identificar la opción correcta dentro de un flujo de diálogo requiriendo razonamiento temporal y aritmético en algunos casos. Mediante un modelo generativo se llego al 75 % de acierto sobre cuatro opciones. En (Khot et al., 2020) las preguntas requieren la composición no evidente de hechos anotados en un corpus de gran tamaño. Aun así, el modelo BERT pasa de un 40 % a un 60 % de acierto cuando se aplica *fine tuning*. Para evitar el sobre-ajuste, los autores refinaron el conjunto de datos añadiendo nuevas respuestas candidatas mediante sistemas adversarios, es decir, entrenando otro modelo para localizar respuestas candidatas distractoras, reduciendo la tasa de aciertos en un 20 %. Otro mecanismo para asegurar la competencia de razonamiento consiste en depurar conjuntos de test seleccionando aquellos en los que los sistemas no superan a los anotadores humanos.

En (Suzgun et al., 2023) se depura BIG-Tech (con 200 tareas) hasta llegar a un conjunto de 23 tareas de resolución de cuestionarios en donde hay acuerdo entre humanos pero los sistemas no obtienen buenos resultados. Los autores muestran que ofreciendo a los sistemas muestras de cadenas de razonamiento mejora los resultados. Esto puede interpretarse como un modo de *forzar* a los modelos de lenguaje a establecer relaciones de implicación a partir de representaciones distribucionales.

Otros marcos de evaluación consisten en preguntas de razonamiento lógico pero no de conocimiento general, sino que se derivan de un pasaje. Es lo que se denomina *Reading Comprehension.* La necesidad de aplicar razonamiento lógico a partir de un pasaje reduce la posibilidad de obtener respuestas correctas a partir de conocimiento de pre-entrenamiento.

Tareas generativas La evaluación de la competencia de razonamiento en tareas generativas es aun un reto para la comunidad. Uno de los primeros conjuntos de test en esta línea es LogicInference, donde se establecen preguntas sobre inferencia lógica combinando lenguaje natural y formalismos lógicos. La respuesta consiste en un texto en lenguaje natural que resuelva el problema lógico planteado. Hasta el momento se ha evaluado esta tarea en términos de ajuste exacto entre la salida del sistema y la respuesta de referencia. Los sistemas mostraron una tasa de acierto bastante limitada (Ontañón et al., 2022). En el conjunto de test e-CARE, se adjunta a la respuesta a un cuestionario una explicación que justifica el razonamiento lógico que lleva a dicha respuesta. La salida del sistema se evalúa mediante una métrica que modela la *fuerza causal* entre palabras a partir de estadísticas de ocurrencias de palabras en en un corpus. La mejor puntuación en los experimentos fue obtenida por GPT-2 con un 0.104 frente al 0.144 obtenido por humanos y un 0.024 obtenido por una aproximación menos efectiva. Los resultados sugieren por tanto que existe en este problema un margen de mejora importante (Du et al., 2022; Dziri et al., 2023).

Otro escenario en el que es necesario restringir los significados de los textos es el acceso a base de datos textuales mediante consultas SQL (Saeed et al., 2023). Este problema tiene una estrecha relación con la población de bases de datos a partir de colecciones de texto. Para realizar esta tarea, es necesario reinterpretar los contenidos textuales de forma que adquieran estructuras y relaciones expresables en lenguaje SQL. Por lo general, este tipo de conjuntos de test se centra en conocimiento enciclopédico. En cuanto a métricas de evaluación, se pueden aplicar en este escenario métricas sencillas (accuracy) sobre resultados de la consulta.

En cualquier caso, la evaluación de la capacidad de razonamiento a partir de una salida generativa es aun un reto por superar, dada la amplia variación lingüística posible entre la salida del sistema y la referencia.

Sin embargo, resulta más sencillo evaluar tareas de razonamiento generativas de tipo matemático, dado que las respuestas deben de ser concisas. Sin embargo, los estudios realizados hasta el momento muestran que las capacidades aritméticas de un modelo de lenguaje son muy limitadas, sobre todo cuando crece el número de símbolos (Qian et al., 2023).

Evaluación de competencias en ODESIA

En esta segunda iteración del proyecto, hemos comenzado a evaluar las competencias generales de los grandes modelos de lenguaje generativos, incorporando un dataset de preguntas de respuesta múltiple correspondientes a once asignaturas de acceso a estudios universitarios (UNED ACCESO), esto es, variación lingüística, composición de contenidos y razonamiento. En este informe de segundo año se incluye la evaluación de varios modelos abiertos y propietarios (GPT4, GPT3.5 y Mistral). Además se está trabajando en la categorización manual de cada pregunta en función de las competencias cognitivas necesarias para contestarlas con vistas a la siguiente iteración del proyecto.

Por otro lado, el corpus de *question answering* con anotación de secuencias SQUAD/SQAC 2024 incluye respuestas que se encuentran distribuidas en diferentes fragmentos.

En general, con el fin de asegurar que las tareas sobre las que se mide la brecha en ODESIA no se limiten a problemas de memorización (variación lingüística) se ha realizado un gran esfuerzo para garantizar que las evaluaciones en ODESIA no estén afectadas por problemas de contaminación. En este sentido, 11 de los datasets utilizados han sido desarrollados en el ámbito del proyecto y no se han compartido las particiones de test de forma pública.

3.3. Informatividad y respuestas engañosas

En secciones anteriores hemos analizado el problema de la evaluación en términos de correctitud de resultados y de la perspectiva de las competencias internas de un modelo de lenguaje. En esta sección analizamos la evaluación de modelos de lenguaje desde la *esperabilidad* de las respuestas. La esperabilidad puede entenderse como la probabilidad de la secuencia devuelta por el sistema dada una distribución de referencia. Por ejemplo, tomando como referencia el corpus sobre el que ha sido pre-entrenado un modelo de lenguaje, la esperabilidad de la secuencia de términos devuelto por el modelo tendría que ver con la probabilidad de la secuencia en dicho corpus. La esperabilidad es por tanto un concepto relativo.

Por definición los sistemas de aprendizaje automático se fundamentan en la maximización de la esperabilidad de la respuesta en el corpus de entrenamiento dadas las restricciones impuestas en la entrada. Por tanto, el aprendizaje automático se basa en la asumción de correspondencia entre esperabilidad y correctitud. Por ejemplo, en el caso de un clasificador entrenado, la respuesta más probable ante una entrada es la respuesta asociada a la instancia del corpus de entrenamiento más similar a dicha entrada.

De esta asunción entre esperabilidad y correctitud surgen dos problemáticas. En primer lugar, surge la pregunta de si un sistema basado en aprendizaje automático es capaz de ofrecer respuestas correctas no esperables. Esto es lo que llamaremos *informatividad* y está relacionado con la capacidad de los clasificadores de acertar en categorías infrecuentes, con la capacidad de los buscadores de ofrecer documentos relevantes poco esperables dada la consulta del usuario, o con a la creatividad de un modelo generativo. La otra cara de la moneda son las respuestas altamente esperables pero incorrectas, lo que lleva a resultados engañosos para el usuario. Esto tiene que ver con los clasificadores que devuelven por sistema la clase mayoritaria, los buscadores y recomendadores basados en popularidad o los modelos generativos que *alucinan*. La alucinación en modelos generativos es la producción de respuestas sin sentido, inconsistentes o incorrectas, pero que en una lectura diagonal pueden resultar creíbles. Esta credibilidad se debe a la maximización de la probabilidad o esperabilidad en términos de cadenas de palabras.

El problema de la esperabilidad ha estado presente desde los comienzos del aprendizaje automático, y ha sido entendido como *sobre-ajuste* a los datos de entrenamiento. Por definición, cualquier sistema basado en aprendizaje devuelve la respuesta más probable dada la entrada y los datos de entrenamiento. La respuesta más probable en los datos de entrenamiento no es siempre correcta, y en muchos casos la respuesta correcta no esté lo suficientemente representada en los datos de entrenamiento. El problema de la alucinación en modelos de lenguaje sugiere que la disponibilidad de inmensas cantidades de datos de pre-entrenamiento no soluciona completamente este problema.

Muy pocos trabajos han planteado la informatividad y la alucinación o engaño como las dos caras de

la misma moneda. Hay algunas excepciones. Lee estudiaron formalmente el equilibrio entre creatividad y alucinación en modelos de lenguaje. Los autores modelan este equilibrio en términos de *likelihood* o esperabilidad en las respuestas.

3.3.1. Informatividad

Entendemos informatividad como la capacidad del sistema de devolver respuestas correctas de baja probabilidad. La informatividad ha sido un aspecto importante de la evaluación en todo tipo de tareas. Por ejemplo, en el caso de los problemas de clasificación, cuando las clases están desequilibradas, es decir, con una clase mayoritaria muy prevalente, se puede obtener una alta tasa de aciertos sin predecir ninguna de las clases minoritarias (Hossin and M.N, 2015). El desequilibrio aparece en muchas aplicaciones en las que la clase positiva se da con una frecuencia reducida, incluidos los datos que se encuentran en el diagnóstico de enfermedades, la detección de fraudes, la seguridad informática y el reconocimiento de imágenes (Johnson and Khoshgoftaar, 2019). La tasa de aciertos no permite medir la capacidad de capturar clases infrecuentes. Para evaluar la informatividad de los clasificadores, se puede evaluar cada clase de manera independiente, para luego hacer un promedio dando el mismo peso a todas ellas. Este es el caso de *Macro Average Accuracy* o la medida F (precisión y recall) aplicado sobre cada clase. Otra opción es aplicar métricas como el coeficiente Kappa, que normaliza la puntuación en función de la distribución de clases en el corpus de test y en la salida del sistema. Otras métricas como el Information Contrast Model puntúan cada acierto o fallo en términos de la cantidad de información dada por la frecuencia de las clases (Amigo and Delgado, 2022). En general, todas estas métricas premian aciertos o penalizan fallos en clases minoritarias.

Algo similar ocurre en problemas de agrupación o *clustering*. Métricas basadas en teoría de la información como *cluster entropy* (Steinbach et al., 2000; Ghosh, 2003), o información mutua (Strehl and Ghosh, 2002) tienen en cuenta la especificidad de los grupos (tamaño) a la hora de premiar o penalizar agrupaciones. Esto hace que, por ejemplo, no se asigne una alta puntuación a agrupar todos los elementos en un único conjunto cuando la distribución de grupos reales está muy desbalanceada.

En problemas de acceso a la información (ordenar productos o documentos por relevancia) la informatividad puede estudiarse de diferentes formas dependiendo del nivel de granularidad respecto al que se mida la especificidad de los documentos o productos a los que se da acceso. Si consideramos la especificidad a nivel de consulta, métricas orientadas a cobertura como Recall at N o NDCG premian el acceso a documentos relevantes en función de la cantidad de documentos relevantes para la consulta en la colección. La métrica *Observational Information Effectiveness* (OIE) tiene en cuenta tanto la cantidad de documentos recuperados por el sistema como la especificidad de la consulta o número de documentos relevantes (Amigó et al., 2022). En otros casos se mide la especificidad de los documentos en función del tema o aspecto abordado por la consulta. Las métricas de diversidad evalúan no solo la relevancia de los documentos sino además su diversidad. Al igual que en clasificación se aplican métricas sobre cada una de las clases, el método más común de capturar la informatividad en recuperación de información consiste en aplicar métricas tradicionales sobre cada uno de los aspectos de la consulta. Este tipo de métricas se denominan *Intent Aware Metrics* (Agrawal et al., 2009). Por último, a nivel de documento o producto individual, la informatividad está íntimamente relacionado con el concepto de *serendipia*, definido como la medida en que un producto que el usuario no esperaba encontrar satisface sus necesidades. Algunas métricas empleadas en este ámbito son la distancia a los productos recomendados por un modelo primitivo que se asume como predecible (Murakami et al., 2008; Ge et al., 2010) o la distancia a los ítems del histórico del usuario (Zuva and Zuva, 2017).

En el contexto de los sistema generativos (generación de lenguaje natural), la informatividad está íntimamente relacionada con el concepto de *creatividad*. Según la real academia española, creatividad es la capacidad de crear nuevas ideas o conceptos, de nuevas asociaciones entre ideas y conceptos conocidos. Existen muchas otras definiciones. Por ejemplo, en el contexto de las ciencias cognitivas encontramos definiciones del tipo *"la combinación y reorganización de elementos, ideas o conceptos preexistentes para dar lugar a algo nuevo y valioso"*, o *"habilidad para percibir conexiones y patrones ocultos en la información, lo que lleva a nuevas ideas y soluciones"*. El contexto de estudio del comportamiento

encontramos *"la capacidad de encontrar soluciones únicas y efectivas para problemas o desafíos"*, en el contexto empresarial *"la creatividad se asocia con la generación de ideas innovadoras que conducen a nuevos productos, servicios o procesos que mejoran la competitividad de una empresa"*, o en el contexto artístico *"capacidad de transmitir emociones, conceptos o narrativas de una manera única que puede resonar con los espectadores o audiencia"*. En cualquier caso, hay dos elementos comunes a cualquier definición de creatividad. Uno es la novedad o grado de sorpresa y el otro es la efectividad o utilidad. En decir, en términos de informatividad, podemos generalizar la noción de creatividad como la capacidad de generar respuestas útiles o correctas pero poco esperables.

En (Franceschelli and Musolesi, 2023) se analiza el desarrollo de los modelos de lenguaje bajo el prisma de las teorías de la creatividad, investigando las cuestiones abiertas y los retos clave. En (Chen and Ding, 2023) se propone una medida de la creatividad consistente en la generación de palabras no relacionadas y el cálculo la distancia semántica entre ellas. Desde este punto de vista, los autores observaron una alta creatividad en modelos como GPT-4, así como un compromiso entre creatividad y estabilidad.

Una opción para evaluar la creatividad de un modelo de lenguaje es tomar como referencia asesores humanos. En (Marco et al., 2022) se presentó a los asesores sinopsis de películas generadas por humanos y por modelos de lenguaje y se preguntó a éstos por el grado de creatividad. En (Summers-Stay et al., 2023) se define una tarea de usos creativos y se evalúan los sistemas mediante juicios humanos preguntando por la *utilidad* y *originalidad* de las respuestas. Se evaluaron diferentes prompts sobre GPT-3. Algunas variantes superaron a los humanos en utilidad y otras en originalidad, pero ninguna de ellas en ambas cosas a la vez. Esto sugiere de nuevo un compromiso entre creatividad y corrección de las respuestas. En (Chakrabarty et al., 2023) se aplica una metodología de evaluación extrínseca estudiando al interacción con un usuarios en una plataforma de asistencia a la escritura creativa.

Además del uso de juicios humanos, ¿tiene sentido definir una métrica de creatividad reutilizable para optimizar sistemas generativos? El problema es cómo medir lo *improbable* de la respuesta. Para ello, es necesario tomar como referencia una espacio de probabilidad. En contextos no computacionales se considera como referencia lo estándar o lo común. Es necesario distinguir entre la capacidad creativa y la creatividad de un objeto o resultado. Por ejemplo, una obra surrealista al puro estilo de Dalí no es creativa si se pintara en los años ochenta, pero su pintor tendría capacidad creativa si éste no hubiese conocido ninguna obra del siglo XX. En el contexto de los modelos de lenguaje, podemos evaluar su capacidad creativa en base a lo que el propio sistema conoce. Es decir, un resultado creativo será aquel que sea efectivo pero no probable según el propio modelo pre-entrenado. Se puede pensar que esto no es posible dado que un modelo de lenguaje siempre devuelve lo más probable según el propio modelo. Sin embargo, hay aspectos que pueden hacer que el modelo se comporte de manera creativa. En primer lugar, el propio proceso generativo (elección de la palabra con mayor probabilidad condicionada al texto que le precede) supone un recorrido en profundidad que puede dar lugar a secuencias de texto improbables. Por ejemplo, partiendo de una palabra probable, la secuencia de n palabras con mayor probabilidad condicionada respecto a las anteriores puede ser en conjunto más impredecible que una palabra de baja probabilidad seguida de n palabras redundantes. Además, mecanismos añadidos como el *prompting* o filtros pueden hacer que los sistemas devuelvan respuestas menos esperables incluso para el propio modelo de lenguaje.

En definitiva, la noción de creatividad y su cuantificación es aun un problema no resuelto. En cualquier caso, el análisis de la literatura sugiere que es necesario de alguna manera medir el grado de sorpresa de la respuesta frente a las fuentes o conocimiento del que parte el sistema.

Evaluación de informatividad en ODESIA

En relación a la informatividad, en la segunda iteración de ODESIA se incorporan métricas que tiene en cuenta la especificidad de las clases en el caso de tareas de clasificación (DIPROMATS y EXISTS). En el ámbito de la experiencia de usuario, la informatividad se evalúa en ODESIA mediante cuestionarios sobre aplicaciones en tareas de acceso a la información (buscadores), tareas de anotación (sistemas de reputación), y de generación de texto (traductores, asistentes virtuales y teclados predictivos).

3.3.2. Resultados Engañosos

En la mayoría de escenarios de desarrollo de sistemas inteligentes, puede asumirse un cierto grado de error, siempre que el usuario pueda identificar dichos errores. Por ejemplo, el usuario de un buscador puede descartar fácilmente documentos no relevantes recuperados por el sistema si la temática difiere sustancialmente de la consulta. Sin embargo, descartar documentos aparentemente relevantes conlleva un mayor coste. En ciertos escenarios críticos como diagnósticos médicos o detección de alarmas en sistemas de reputación, un falso positivo no identificable por el usuario puede suponer un alto coste. Hoy en día, las respuestas engañosas son el motivo principal por el que descartar el uso de sistemas inteligentes en escenarios críticos, por encima de la tasa de error.

En la sección anterior, definimos la informatividad como la capacidad de generar respuestas correctas improbables. Lo opuesto a esto es *la generación de respuestas probables o esperables pero incorrectas*. Dado que en general los sistemas de aprendizaje automático maximizan la probabilidad de la salida, éste es el tipo de respuesta erróneas más común. Una respuesta incorrecta esperable puede resultar engañosa.

En tareas de clasificación con clases no balanceadas los sistemas tienden a seleccionar la clase mayoritaria con el fin de maximizar la probabilidad de la respuesta. Un falso positivo sobre una clase común es difícil de identificar, por lo que suele resultar engañoso para el usuario. Por ejemplo, en un escenario de detección correos spam, en donde la mayoría de los mensajes no lo son, es más fácil para el usuario identificar un falso spam que identificar un spam en la bandeja de entrada. En un escenario de identificación de alarmas (por ejemplo en sistemas de monitorización de reputación en redes sociales, detección de fraude, etc.) es más sencillo identificar una falsa alarma que una alarma oculta entre los datos de entrada. Al igual que en el caso de la informatividad analizada en la sección anterior, el sesgo hacia la clase mayoritaria se evalúa mediante métricas tradicionales aplicadas sobre cada una de las clases. Para evaluar este mismo aspecto en sistemas de agrupación se aplican métricas basadas en teoría de la información que acentúan el efecto de las decisiones sobre grupos minoritarios. En sistemas de acceso a la información (ranking de documentos y productos) también se aplican los mismos mecanismos de evaluación que en el problema de la informatividad (ver sección anterior).

Sin embargo, con el auge de los sistemas generativos basados en pre-entrenamiento, la informatividad y la generación de respuestas engañosas debe evaluarse de manera independiente. En concreto, las respuestas engañosas en modelos generativos está íntimamente relacionado con el problema de la alucinación. La alucinación en sistemas generativos ha sido definida por diversos autores como *"contenido generado que no tiene sentido o que no se corresponde con contenido de las fuentes"* (Ji et al., 2023). Esto incluye tanto contenidos contradictorios respecto a las fuentes como contenidos que no pueden ser verificados. Es lo que Ji et al. definen como alucinación intrínseca y extrínseca. Sin embargo, implícitamente se asume que la alucinación son contenidos que en un momento determinado podrían ser interpretados como ciertos por el usuario. Por ejemplo, Filippova definen la alucinación como textos fluidos pero no soportados por las fuentes (Filippova, 2020). Es importante distinguir la alucinación de otros fenómenos relacionados. Por ejemplo, la *factualidad* se refiere a información consistente con hechos reales (Maynez et al., 2020). También puede darse casos en los que el sistema genera contenidos falsos procedentes de bulos o creencias falsas presentes en la colección de pre-entrenamiento. En (Lin et al., 2022) se desarrolla un conjunto de test con este propósito. En este documento definimos la alucinación como **el resultado de generar respuestas de incorrectas de alta probabilidad en modelos generativos**. Esto implica respuestas fluidas, aparentemente correctas, pero inconsistentes con las fuentes o simplemente no verificables.

El problema de la evaluación de este problema está aun en discusión en la comunidad. La característica común de todas las soluciones es que se contrasta la respuesta del sistema con las fuentes mediante diferentes técnicas.

Proximidad textual: Una primera aproximación es comparar a nivel textual la respuesta con las fuentes. Dada la variación lingüística entre respuestas y fuentes, esto se ha aplicado cuando las fuentes tienen contenido semi-estructurado como tablas. Este es el caso de la métrica PARENT (Dhingra et al., 2019). La métrica *knowledge F1* aplica solapamiento de palabras con las fuentes pero centrándose en los textos empleados por los asesores en la anotación para evitar el efecto de la variación lingüística

(Shuster et al., 2021).

Extracción de información: Se extrae conocimiento estructurado de la respuesta y de las fuentes mediante herramientas de extracción de información y se comparan entre sí. Por ejemplo, en (Goodrich et al., 2019) se extraen tripletes del tipo sujeto-relación-objeto en el contexto del resumen abstractivo. La limitación de este método es la acumulación de errores en la fase de extracción de información.

Búsqueda de respuesta: Otros autores proponen el uso de herramientas de búsqueda de respuestas. Se generan automáticamente preguntas a partir del texto de referencia y se aplica un método de búsqueda de respuestas sobre las fuentes y sobre el texto generado. Las respuestas obtenidas deberían de ser consistentes. Esto ha sido aplicado en problemas de resumen abstractivo (Durmus et al., 2020). En el contexto de sistemas de diálogo Honovich et al. introducen herramientas de detección de inferencia para comparar las respuestas (Honovich et al., 2021).

Inferencia textual: En muchos contextos se ha aplicado sistemas de detección de implicación textual entre las fuentes y el texto generado. Esto ha sido aplicado en diversas tareas como sistemas de diálogo (Dziri et al., 2022), resumen abstractivo (Huang et al., 2021; Kryscinski et al., 2020b).

Desde una perspectiva global, todas estas métricas tienen dos inconvenientes. El primero es que detectar automáticamente la consistencia entre las fuentes y el texto generado es tan complicado, y por la tanto impreciso, como la tarea en si de generación de texto. Por tanto, la fase de evaluación puede estar sesgada por las técnicas empleadas. El segundo inconveniente es que inevitablemente se penaliza la creatividad. Cuando más aporte el texto generado respecto a las fuentes, menos encajará éste con las fuentes.

Evaluación de contenidos engañosos en ODESIA:

En relación a respuestas engañosas, en ODESIA se incorporan métricas que tiene en cuenta la especificidad de las clases en el caso de tareas de clasificación con alto grado de desequilibrio, es decir, donde la tasa de aciertos sobre una clase mayoritaria puede resultar engañosa para el usuario. En el ámbito de la experiencia de usuario, las respuestas engañosas se evalúan en ODESIA mediante cuestionarios sobre aplicaciones en tareas de acceso a la información (buscadores), tareas de anotación (sistemas de reputación), y de generación de texto (traductores, asistentes virtuales y teclados predictivos).

4. Definición de indicadores

En esta segunda iteración del proyecto se mantiene la misma definición de indicadores que en el primer año. Incluimos en esta sección la definición de estos indicadores con el fin de que el documento sea autocontenido.

4.1. Indicadores Ámbito 1: Estado del Arte

4.1.1. Indicadores de diseminación

Los indicadores de diseminación en ODESIA miden el estado de las tecnologías en español e inglés en cuanto a la divulgación de resultados en medios científicos. Para el cálculo de estos indicadores se tienen en cuenta los artículos publicados y los proyectos subvencionados. En cuanto a los artículos publicados (Indicador D.1), las publicaciones se buscan en los proceedings de congresos recientes de PLN de índole internacional. Para determinar si un artículo o proyecto está orientado al español, inglés o a ambos se determina si la experimentación se ha realizado sobre datos en español, inglés o en ambas lenguas. La identificación de estos artículos se realiza mediante una combinación de búsquedas por palabras clave y revisión manual (en el primer año, se realizó un estudio en profundidad de 50 artículos para comprobar el margen de error del proceso). En cuanto a proyectos subvencionados, se analizan proyectos a nivel europeo y de Estados Unidos. En la secciones correspondientes a resultados se proporcionan más detalles sobre el proceso de recopilación de datos en esta segunda iteración.

El indicador de publicaciones se calcula como la diferencia entre publicaciones que trabajan con textos en inglés y publicaciones que trabajan con textos en español dividido por la suma de ambos. La brecha obtenida sería nula solo en el caso de que el número de publicaciones en inglés sea el mismo que el número de publicaciones que cubran además en español. La brecha sería negativa y del 100 % si todos los artículos cubrieran el español y ninguno el inglés. Sería positiva del 100 % si ningún artículo cubriera el español y todos ellos cubrieran el inglés.

Indicador D.1: Brecha en publicaciones

Representa la diferencia porcentual en cuanto a publicaciones en tecnologías de la lengua en castellano. Se obtendrá de la siguiente forma:

$$\text{D. } 1 = \frac{|P_I| - |P_E|}{|P_E \cup P_I|} \cdot 100$$

donde $|P_E|$ y $|P_I|$ representan el número de publicaciones de tecnologías que cubren el español y que cubren exclusivamente el inglés.

En relación al indicador D.2 (Brecha en proyectos subvencionados), se realizan búsquedas en bases de datos públicas de proyectos de investigación subvencionados. En las búsquedas se usan filtros disponibles para seleccionar proyectos de PLN. Posteriormente, se filtran los proyectos obtenidos en la primera búsqueda conforme al intervalo temporal considerado. A continuación se revisa manualmente el título y la descripción de los proyectos, para hallar los que experimentan sobre datos en español y los que experimentan sobre datos en inglés.

El indicador se calculará como la diferencia entre proyectos exclusivamente en inglés y proyectos que además cubran el español dividido por la suma de ambos. La brecha obtenida sería nula solo en el caso de que el número de proyectos exclusivamente en inglés sea el mismo que el número de proyectos que cubran además el español. La brecha sería negativa y del 100 % si todos los proyectos cubrieran además el español. Sería positiva del 100 % si ningún proyecto cubriera el español.

Indicador D.2: Brecha en proyectos subvencionados.

Representa la diferencia porcentual en cuanto a proyectos subvencionados en tecnologías de la lengua en castellano. Se obtendrá de la siguiente forma:

$$\text{D. } 2 = \frac{|P_I| - |P_E|}{|P_E \cup P_I|} \cdot 100$$

donde $|P_E|$ representa proyectos donde se cubre el español y $|P_I|$ representan el número de proyectos de tecnologías exclusivamente en inglés.

4.1.2. Indicadores de recursos

Consideraremos como *recursos* en tecnologías de la lengua los corpora de texto disponible en ambos idiomas, y los modelos de lenguaje que se usen en el desarrollo de tecnologías de PLN.

4.1.3. Indicador de texto disponible en internet

Los modelos de lenguaje generativos que se han desarrollado en los últimos años y que lideran los avances en la inteligencia artificial relacionada con el lenguaje están entrenados con grandes masas de texto disponibles en Internet. Este indicador medirá la brecha entre el español y el inglés en cuanto a a disponibilidad de texto en Internet para estas lenguas. Para calcularlo se emplearán la información para inglés y español:

- Número de artículos en Wikipedia.
- Porcentaje de páginas en Internet (tomando como referencia el número de páginas para las que se conoce la lengua).
- Número de textos en Internet Archive.
- Número de textos en PubMed.
- Porcentaje de páginas en el último crawl de Common Crawl.

Indicador R.0: Brecha en masa de texto disponible en internet

Representa la diferencia porcentual en cuanto a disponibilidad de corpus de texto en las respectivas lenguas. Sea F el conjunto de fuentes (wikipedia, etc.), y sea P_f el peso asignado a fuente:

$$\text{R. } 0 = \sum_{f \in F} P_f \frac{T_I^f - T_E^f}{T_I^f + T_E^f} \cdot 100$$

donde T_I^f y $\frac{f}{E}$ representan el volumen de texto en la fuente f en inglés y en español respectivamente.

4.1.4. Indicadores de modelos de lenguaje pre-entrenados

Los últimos avances en tecnología de la lengua muestran sistemáticamente que disponer de modelos pre-entrenados se traduce en una mejora significativa en términos de efectividad en la gran mayoría de problemas.

Para ello, explotamos la información disponible en Hugging Face,[6] la biblioteca de código abierto más popular actualmente, que permite a los desarrolladores utilizar modelos pre-entrenados para tareas como clasificación de texto, traducción automática y generación de texto, entre otros. Los modelos se organizan en categorías basadas en el tipo de tarea de procesamiento que abordan, como la clasificación de texto, la generación de texto y la traducción automática. Estas tareas se corresponden de forma aproximada con la categorización de aplicaciones desde un punto de vista técnico descrita en la sección 2.3. Además, Hugging Face permite filtrar por idioma, arquitectura de modelo y otros criterios relevantes.

Indicador R.1: Brecha en modelos de lenguaje

Representa la diferencia porcentual en cuanto a disponibilidad de modelos de lenguaje. Sea D el conjunto de tareas o dominios considerados, y sea P_d el peso asignado a cada tarea o dominio:

$$\text{R. } 1 = \sum_{d \in D} P_d \frac{|M_I| - |M_E|}{|M_I \cup M_E|} \cdot 100$$

donde M_I y M_E representan los conjuntos de modelos entrenados sobre texto en inglés y en español respectivamente. Los modelos multilingües seran considerados como pertenecientes a ambos conjuntos.

4.1.5. Indicadores de datos anotados

En muchos escenarios, el éxito de los sistemas está condicionado por la disponibilidad de datos de entrenamiento, especialmente en cierto tipo de tareas de minería de textos (ver Tabla **??**) en donde la información implícita que debe ser extraída de los textos depende del dominio de aplicación y del problema concreto. Algunos ejemplos son la clasificación temática de textos en dominios específicos, el análisis de sentimientos, o tareas de extracción de información. En estos casos no es suficiente pre-entrenar los modelos directamente sobre grandes colecciones de documentos, sino que es necesario disponer de datos anotados manualmente específicos para cada problema.

Podemos distinguir entre tres tipos de datos anotados. En primer lugar, aquellos que son anotados por expertos en un laboratorio. El número de datos anotados es limitado dado que la contratación de expertos es cara. En segundo lugar, estos costes pueden reducirse mediante el uso de servicios de anotación on-line de no expertos. La calidad de la anotación es menor, pero puede compensarse con el análisis de los datos, descartando anotaciones incoherentes o incluyendo pruebas de comprobación en los propios datos a anotar. Una tercera vía consiste en explorar información colaborativa, es decir, información anotada por los propios usuarios durante el uso de sistemas. Por ejemplo, los *logs* de navegación son una herramienta clave para entrenar motores de búsqueda. Es posible también emular datos anotados de análisis de sentimiento a partir de emoticonos y feedback de los usuarios.

Para la recopilación de datos se consideran campañas de evaluación competitivas y repositorios de recursos lingüísticos. Concretamente, se estudiará la presencia de datos anotados en ambas lenguas en las principales campañas de evaluación y repositorios a nivel nacional, europeo e internacional. Generalmente, las campañas de evaluación como CLEF, IBERLEF o SEMEVAL aglutinan una serie de tareas en las que

[6] https://huggingface.co/

grupos de investigación presentan sus aproximaciones a problemas específicos que son evaluadas sobre un conjunto de datos anotados común para cada tarea.

La brecha se computa como los recursos encontrados para el inglés menos los encontrados para el castellano dividido por la suma de ambos. La brecha será nula si los recursos en ambas lenguas son igualmente numerosos, obteniendo una brecha positiva o negativa del 100 % si solo hubiera recursos en uno de los idiomas. Los criterios de ponderación se especificarán en el informe de resultados de cada iteración.

Indicador R.2.a: Presencia de datos anotados en repositorios

Representa la diferencia porcentual en cuanto a disponibilidad de corpus anotado para entrenamiento y test en foros internacionales (campañas de evaluación competitiva y repositorios). Sea R el conjunto de repositorios y P_r el peso asignado a cada repositorio:

$$\text{R. 2. a} = \sum_{r \in \mathcal{R}} P_r \frac{|D_I| - |D_E|}{|D_I \bigcup D_E|} \cdot 100$$

donde D_I y D_E representan los conjuntos de datos anotados en inglés y castellano respectivamente en repositorios públicos.

Indicador R.2.b: Presencia de datos anotados en campañas de evaluación

Representa la diferencia porcentual en cuanto a disponibilidad de corpus anotado para entrenamiento y test campañas de evaluación competitiva. Sea $\mathcal{C}$ el conjunto de campañas de evaluación y P_c el peso asignado a cada campaña:

$$\text{R. 2. b} = \sum_{c \in \mathcal{C}} P_c \frac{|T_I| - |T_E|}{|T_I \bigcup T_E|} \cdot 100$$

donde T_I y T_E representan el conjunto de tareas competitivas en las campañas con datos anotados en inglés y castellano respectivamente.

El tercer indicador de recursos anotados tiene como objetivo obtener una visión general de aquellos dominios y tareas en los que se dispone de datos anotados para el español, identificando a su vez escenarios en los que existe aún un vacío. Además, se considerarán las tareas de procesamiento de lenguaje (lematización, análisis de dependencias, reconocimiento de entidades nombradas etc.) como un dominio adicional.

Indicador R.2.c: Brecha en diversidad de datos anotados

Representa la diferencia porcentual en cuanto a disponibilidad de corpus anotado para entrenamiento y test en diferentes aplicaciones. Sea $\mathcal{D}$ el conjunto de dominios, P_d el peso asignado a cada dominio, y $\mathcal{H}_d$ el conjunto de familias de aplicaciones identificadas para dicho dominio, se calcula:

$$\text{R. 2. c} = \sum_{d \in \mathcal{D}} P_d \sum_{h \in \mathcal{H}_d} \frac{C_I^h - C_E^h}{C_{I \cup E}^h}$$

donde C_I^h y C_E^h toman el valor uno o cero dependiendo de si existen datos anotados para la familia de aplicaciones h en el idioma correspondiente. $C_{I \cup E}^h$ toma valor 1 si existen datos anotados en cualquiera de los dos idiomas. Solo se considerarán familias de aplicaciones con datos en al menos uno de los dos idiomas.

Indicador R.3: Brecha en efectividad de modelos para tareas de procesamiento de lenguaje

Representa la diferencia entre idiomas entre mejoras porcentuales efectividad de herramientas de procesamiento sobre un sistema base no lingüístico. Sea $\mathcal{H}$ el conjunto de herramientas seleccionadas, y P_h el peso asignado a cada herramienta, se calcula:

$$\text{R. 3} = \sum_{h \in \mathcal{H}} P_h \left(\frac{s_I^h - b_I^h}{|b_I^h - r^h|} - \frac{s_E^h - b_E^h}{|b_E^h - r_0^h|} \right)$$

donde s_I^h, b_I^h y r^h representan la efectividad de la herramienta h, del sistema base y el valor de efectividad de referencia en inglés. La notación para el español es análoga.

4.1.6. Indicadores de efectividad

Comúnmente, las aplicaciones en tecnologías de la lengua se evalúan en base a criterios extrínsecos, es decir, en base a la eficacia del sistema en tareas específicas. Estas metodologías están relacionadas con la usabilidad y captan la capacidad de aprender y generalizar en el contexto de problemas específicos. Estos marcos de evaluación pretenden capturar escenarios de uso como la traducción automática, la respuesta a

preguntas, el resumen automático, la minería de opiniones, etc. Idealmente, los datos de prueba deberían ser una muestra representativa del escenario de uso del sistema. La fortaleza de estos marcos de evaluación radica en que proporcionan información directa sobre la usabilidad del sistema. El punto débil es el efecto del sobre-ajuste. No siempre es fácil recoger los datos de entrada-salida en un escenario real. Por ejemplo, no hay demasiados datos de entrenamiento disponibles para las lenguas poco comunes en la traducción automática o para los temas dominios no populares en sistemas de diálogo.

Definir un indicador de brecha en efectividad entre lenguas requiere tener en cuenta muchas variables, que en muchos casos dependen también del problema y de los datos disponibles para la evaluación. El primer problema a resolver es que no todas las métricas de efectividad tienen las mismas propiedades de escala. La mayoría de las métricas, como por ejemplo la tasa de aciertos, están acotadas entre cero y uno. Otras métricas no tienen una cota superior. Por tanto, no se pueden equiparar las diferencias obtenidas para varios problemas en los que se emplean diferentes métricas. Es decir, un intervalo en una métrica puede tener una relevancia completamente distinta del mismo intervalo en otra métrica. Por tanto, es necesario establecer un intervalo-unidad o intervalo de referencia.

El segundo gran problema es que la efectividad obtenida puede ser sensible a la dificultad intrínseca de los datos de evaluación. Por ejemplo, sobre un tamaño de datos de entrenamiento menor, es normal obtener valores de efectividad menores. Esto no significa que exista una brecha intrínseca en cuanto a efectividad de los sistemas en cada uno de los idiomas. Para controlar este aspecto, será necesario tomar como referencia un sistema base que cumpla ciertas características. El sistema base no debe incluir ningún tipo de tecnología de la lengua específica de idioma. Es decir, debe de ser un sistema no pre-entrenado para ningún idioma, y que además no emplee herramientas de procesamiento lingüístico. Por ejemplo, emplear un clasificador SVM sobre conjuntos de *tokens* para clasificar textos no supone pre-entrenamiento ni pre-procesamiento lingüístico. Por tanto, las diferencias de efectividad entre idiomas vendrán determinadas únicamente por la dificultad del conjunto de datos de evaluación.

Teniendo en cuenta estos dos factores, tomaremos como intervalo de referencia en cada idioma la distancia entre la efectividad del sistema base que denotaremos como b, y un punto de referencia en la escala de la métrica que denotaremos como r. Es decir, tomaremos como intervalo unidad $|b - r|$.

El punto de referencia r también requiere un análisis previo. En casos en los que la métrica no esté acotada superiormente o en casos en los que la efectividad de los sistemas sea muy baja, este punto debería ser la cota inferior. Por otro lado, en casos en los que exista una cota superior y la efectividad sea alta debería tomarse como punto de referencia el valor superior en la escala de la métrica. Por ejemplo, dado un sistema base con efectividad cercana al uno en una métrica acotada superiormente, por ejemplo, 82 % de acierto, y tomando como punto de referencia el 100 % de acierto, el intervalo unidad debería ser del 18 %.

Una vez definido este intervalo unidad $|b-r|$ la aportación lingüística efectiva en un idioma se calculará como el ratio de la diferencia entre efectividad del sistema evaluado y el sistema base respecto al intervalo unidad:

$$\Delta = \frac{s - b}{|b - r|} \cdot 100$$

La aportación lingüística en cada idioma cumple las siguientes propiedades. En primer lugar, la aportación es nula cuando el mejor sistema se comportan igual que el sistema base ausente de tecnología lingüística:

$$s = b \Longrightarrow \Delta = 0$$

En segundo lugar, dada una efectividad fija por parte del sistema base, la aportación es proporcional a la diferencia entre efectividad del sistema evaluado y la del sistema base:

$$b = k \Longrightarrow \Delta \propto s - b$$

En tercer lugar, dado una diferencia fija entre el sistema y el sistema base, la contribución será proporcional a la inversa del intervalo unidad:

$$s - b = k \Longrightarrow \Delta \propto \frac{1}{|b - r|}$$

Esto quiere decir que, tomando como referencia la máxima puntuación ($r = 1$) a medida que la efectividad del sistema y del sistema base se aproximen al punto máximo, la contribución será mayor. Por ejemplo, una mejora de 0.97 a 0.98 es más importante que una mejora de 0.67 a 0.68. De la misma forma, a valores bajos de efectividad, tomando como punto de referencia ($r = 1$), una mejora de 0.1 a 0.2 será más significativa que una mejora de 0.3 a 0.4.

El indicador de la brecha de efectividad entre idiomas inglés (I) y español (E) se calculará mediante el indicador:

$$Ind(I, E) = \Delta_I - \Delta_E = \frac{s_I - b_I}{|b_I - r|} - \frac{s_E - b_E}{|b_E - r|}$$

Este indicador cumple las siguientes propiedades. En primer lugar, es simétrico respecto a los idiomas:

$$Ind(I, E) = -Ind(E, I)$$

En segundo lugar, un comportamiento idéntico en ambos idiomas se corresponde con una brecha cero:

$$Ind(I, I) = Ind(E, E) = 0$$

Tomando como punto de referencia $r = 0$, se obtiene una brecha nula cuando ambos presentan la misma diferencia porcentual respecto al baseline.

$$\left.\begin{array}{l} r = 0 \\ \frac{s_I - b_I}{b_I} = \frac{s_E - b_E}{b_E} \end{array}\right\} \Rightarrow Ind(I, E) = 0$$

que es lo mismo que decir que ambos tienen la misma proporción de efectividad respecto a sus sistemas base.

$$\left.\begin{array}{l} r = 0 \\ \frac{s_I}{b_I} = \frac{s}{b_E} \end{array}\right\} \Rightarrow Ind(I, E) = 0$$

Tomando como punto de referencia $r = 0$, se obtiene una brecha del 100 % cuando la efectividad del sistema en ingles consigue la diferencia porcentual conseguida por el sistema en español (brecha cero) más la efectividad del sistema base en inglés.

$$\left.\begin{array}{l} r = 0 \\ s_I = b_I \cdot \frac{s_E}{b_E} + b_I \end{array}\right\} \Rightarrow Ind(I, E) = 100\,\%$$

Tomando como referencia $r = 1$ (cota máxima), la brecha es nula cuando la efectividad del sistema en ambas lenguas es proporcional la diferencia entre los sistemas base y la cota superior:

$$\left.\begin{array}{l} r = 0 \\ \frac{s_I}{1 - b_I} = \frac{s}{1 - b_E} \end{array}\right\} \Rightarrow Ind(I, E) = 0$$

Tomando como referencia $r = 1$, hay una brecha del 100 % cuando el sistema en español no supera al sistema base, mientras que el sistema en inglés obtiene la máxima puntuación.

$$\left.\begin{array}{l} r = 1 \\ s_I = 1 \\ s_E = b_E \end{array}\right\} \Rightarrow Ind(I, E) = 100\,\%$$

Para este indicador se emplearán dos fuentes de datos. Por un lado, experimentos en laboratorio dentro del contexto del proyecto en el que se compararán para ambos idiomas sistemas base carentes de tecnología linguística con modelos pre-entrenados y re-entrenados sobre datos anotados. Estas fuentes tienen la ventaja de ser experimentación controlada orientada a la brecha lingüística. La limitación es que no será posible cubrir muchas de las familias de aplicaciones. Estos datos se complementarán con el análisis de

contribuciones en la literatura donde se aporten resultados de experimentos en distintas lenguas sobre sistemas base y datos anotados comparables.

Indicador E.1: Brecha en efectividad

Representa la diferencia entre idiomas entre mejoras porcentuales sobre un sistema base no lingüístico. Sea $\mathcal{D}$ el conjunto de dominios, P_d el peso asignado a cada dominio, y $\mathcal{H}_d$ el conjunto de aplicaciones seleccionadas para dicho dominio, se calcula:

$$\text{E. } 1 = \sum_{d \in \mathcal{D}} P_d \sum_{h \in \mathcal{H}_d} \frac{s_I^h - b_I^h}{|b_I^h - r^h|} - \frac{s_E^h - b_E^h}{|b_E^h - r_0^h|}$$

donde s_I^h, b_I^h y r^h representan la efectividad de la herramienta h, del sistema base y el punto de referencia en inglés. La notación para el español es análoga.

4.2. Indicadores Ámbito 2: Soluciones de mercado

En este ámbito se realiza un análisis comparativo de la oferta de productos dentro del campo de las tecnologías de procesamiento del lenguaje natural en inglés y en español. Para calcular la brecha se hacen los siguientes pasos. En primer lugar, se seleccionan los productos y servicios que incorporan tecnologías del lenguaje de cinco áreas representativas y que están disponibles en el mercado. A continuación, se compila un listado completa de las funcionalidades de estas soluciones de mercado. Después, se lleva a cabo una comparativa detallada de las funcionalidades de estas soluciones y, en base a esta comparativa, se ha calculado el indicador de brecha en funcionalidades. Para elegir las funcionalidades en primer lugar se identifican las necesidades del usuario en cada una de las áreas de aplicaciones y a continuación se evalúan las funcionalidades disponibles que satisfacen esas necesidades mediante el uso de tecnologías de la lengua.

4.2.1. Selección de productos y servicios

En consonancia con el estudio realizado en el primer año, se han seleccionado las siguientes áreas de aplicaciones por ser las que tienen un uso más común entre ciudadanos y tener un mayor impacto en la industria: análisis de opiniones, asistentes virtuales, traducción automática, teclados predictivos y buscadores web. Dentro de estas àreas se han seleccionado las aplicaciones que se describen a continuación, un total de 51. Para cada aplicación se indica el criterio seguido para su selección. Las herramientas se distinguen entre aquellas que son ofertadas, principalmente, para el uso de ciudadanos o consumidores finales, y las que son ofertadas para un uso empresarial o profesional. Las herramientas de uso ciudadano aparecen con fondo blanco, mientras que las de uso empresarial con fondo amarillo.

El criterio principal para la selección ha sido la popularidad. En ausencia de una fuente homogénea, se ha medido esta popularidad utilizando el número de usuarios, descargas, reseñas, visitas, comentarios, cuota de mercado, ingresos generados o informes de expertos. En la medida de lo posible, se ha intentado utilizar un único criterio para comparar las soluciones de uso ciudadano y de uso empresarial de cada una de las áreas.

Área de análisis de opiniones. Se han seleccionado 10 soluciones de uso empresarial. Los datos de ingresos se han obtenido de (Little, 2021; ZoomInfo Technologies LLC, 2023).

- Sprinklr — Ingresos: 618.2 M$.
- Khoros — Ingresos: 252.1 M$.
- NetBase Quid — Ingresos: 64.5 M$.
- Brandwatch — Ingresos: 210.1 M$.
- Linkfluence — Adquirida por Melwater. Ingresos: 37.4 M$.
- Synthesio — Ingresos: 25.9 M$.
- Talkwalker – Ingresos: 84 M$.
- Digimind — Ingresos: 42 M$.
- Resonate — Ingresos: 22.8 M$.

- Meltwater — Anteriormente Sysomos. Ingresos: 462.1 M$.

Área de asistentes virtuales. Se han seleccionado 6 herramientas de uso ciudadano y 5 herramientas de uso empresarial (Boost.ai, 2023).

- Google Assistant — más de 2.5 B de usuarios de Android activos en el mundo y 70 % de cuota en el uso de teléfonos inteligentes (Curry, 2023).
- Siri — 19 % de los dispositivos móviles en el mundo (Counterpoint Technology Market Research, 2023).
- Alexa — Más de 100 M de usuarios (Smith, 2023).
- Bixby — Más de 200 M de usuarios (Samsung, 2022).
- Kore.ai — Ingresos: 154 M$ (ZoomInfo Technologies LLC, 2023).
- IBM Watson Assistant — No se ha encontrado información sobre facturación. Segundo en ranking (PeerSpot, 2023).
- Amazon Lex — No se ha encontrado información sobre facturación. Tercero en ranking (PeerSpot, 2023).
- Google Dialogflow — No se ha encontrado información sobre facturación. Primero en ranking (PeerSpot, 2023).
- Amelia — Ingresos: 11.3 M$ (ZoomInfo Technologies LLC, 2023).
- ChatGPT — Más de 100 M de usuarios activos por semana (Shewale, 2023c).
- Google Bard — Más de 140.6 M de visitantes mensuales (Shewale, 2023a).

Área de traducción automática. Se han seleccionado 6 herramientas de uso ciudadano y 6 herramientas de uso empresarial (greatcontent GmbH, 2023) (greatcontent GmbH, 2023), cuyo número de visitas se obtuvo de (Similarweb LTD, 2023).

- Google Translate — 713.5 M visitas al mes.
- DeepL — 246.3 M visitas al mes.
- Bing Translator o Microsoft Translator
- Amazon Translate
- Systran Translate — 413.3 K visitas al mes.
- Reverso Translation — 89.6 M visitas al mes.
- memoQ Translator PRO
- Smartling — 581 K visitas al mes.
- Crowdin — 1.9 M visitas al mes.
- TextUnited — 33.3 K visitas al mes.
- ChatGPT — 1.7 B visitas al mes.
- Google Bard — 266.1 M visitas al mes.

Área de teclados predictivos. Se han seleccionado 7 herramientas de uso ciudadano y 2 herramientas de uso empresarial según los ratings obtenidos en App Store y Google Play, salvo que se indique otra referencia.

- Microsoft SwiftKey — 103 K ratings en App Store y 4.03 M ratings en Google Play.
- GBoard — 44.6 K ratings en App Store y 13.7 M ratings en Google Play.
- Grammarly Keyboard — 95.9 K ratings en App Store y 207 K ratings en Google Play.
- Fleksy — 739 ratings en App Store, 270 K ratings en Google Play.
- iPhone Keyboard — 19 % de los dispositivos móviles en el mundo (Counterpoint Technology Market Research, 2023).
- GMail (funciones predictivas en la redacción) — Más de 1.8 B de usuarios activos (Shewale, 2023b).

- Google Workspace (funciones predictivas en la redacción) — Más de 3.000 M de usuarios y más de 8 M de usuarios de pago (Izatt, 2022).
- Microsoft Outlook (funciones predictivas en la redacción) — 400 M de usuarios (Silva-Payne, 2022).
- Microsoft Office 365 (funciones predictivas en la redacción) — Más de 345 M usuarios de pago (Redmond, 2022).

Área de buscadores web. Se han seleccionado 6 herramientas de uso ciudadano (GmbH, 2023) y 3 herramientas de uso empresarial.

- Google Search — 83,49 % de cuota de mercado.
- Bing — 9.19 % de cuota de mercado.
- Yahoo Search — 2,72 % de cuota de mercado.
- DuckDuckGo — menos de 1 % de cuota de mercado.
- Brave Search — 169.9 M visitas.
- Elasticsearch — líder en software semi-libre (Gartner, Inc., 2023).
- Mindbreeze — líder en software propietario (Gartner, Inc., 2023).
- Apache Solr — representativo de software libre seleccionado por criterio experto de científicos de datos.
- Perplexity — 47.5 M visitas.

4.2.2. Listado de funcionalidades

En esta sección se detallan las funcionalidades que se valorarán para cada una de las áreas. Para elegir las funcionalidades en primer lugar se han identificado las necesidades del usuario en cada una de las áreas de aplicaciones mediante criterio experto y a continuación se han evaluado las funcionalidades disponibles que satisfacen esas necesidades mediante el uso de tecnologías de la lengua inspeccionando los distintos productos.

Aquellas funcionalidades en las que aparece un (+1) o (+2) junto al nombre contienen, bajo esa misma definición, una o dos funcionalidades adicionales respectivamente, elevando el total a dos o tres. Si aparece un (+n) significa que el número total de funcionalidades adicionales bajo esa definición no está limitado.

Análisis de opiniones

- **Clasificación de sentimiento (+1)**: Capacidad de clasificar el mensaje según la polaridad del sentimiento mostrado por su emisor. Se valorará la posibilidad de clasificar los mensajes neutrales como una funcionalidad específica.
- **Clasificación de impacto reputacional (+1)**: Capacidad de clasificar el mensaje según el impacto que tiene sobre la reputación de una determinada entidad. Se valorará la posibilidad de clasificar los mensajes neutrales como una funcionalidad específica
- **Clasificación de emociones (+n)**: Detección del tipo de emoción expresada en un mensaje (rabia, repulsión, felicidad, etc.). Cada una de las clases que la solución detecte contará como una funcionalidad.
- **Detección de tema de conversación**: Detección del tema de conversación.
- **Detección de mensajes inapropiados (+n)**: Detección de mensajes que pueden ser dañinos o no aptos para todos los públicos. Cada una de las categorías detectadas (mensajes de odio, acoso, pornográficos, etc.) contará como una funcionalidad.
- **Detección de entidades**: Capacidad de detectar entidades (nombres de marcas, productos o personas).
- **Detección de motivaciones (psychological drivers)**: Capacidad de detectar motivaciones o estímulos de los emisores de los mensajes, intenciones o posibles hábitos.

- **Detección de sentimiento para cada aspecto de un producto**: Análisis de sentimiento sobre un aspecto o atributo del producto mencionado en el texto o indicado de antemano.
- **Detección de sentimiento por entidades (+1)**: Análisis de sentimiento sobre cada una de las entidades identificadas en el texto. Se valorará la posibilidad de clasificar los mensajes neutrales como una funcionalidad específica.
- **Posibilidad de definir las clases (+n)**: Posibilidad de que el usuario pueda definir las clases a detectar. Podrían ser emociones, polaridad de sentimiento o reputacional, temas, etc.
- **Posibilidad de ajustar el modelo (+n)**: Posibilidad de ajustar cualquiera de los modelos anteriores para que se adapte a una determinada temática o criterio.

Asistentes virtuales

- **Reconocimiento de voz**: Capacidad de detectar la identidad de los usuarios mediante el análisis de patrones de voz únicos y características físicas de su voz.
- **Posibilidad de añadir habilidades (+n)**: Capacidad de interactuar con otros servicios como diarios, aplicaciones. Cada uno de los servicios relevantes detectados contará como una funcionalidad.
- **Acentos regionales (+n)**: Capacidad de comunicarse con vocabularios y acentos regionales. Cada uno de los acentos regionales contará como una funcionalidad.
- **Comandos aceptados (+n)**: Los comandos que entiende el asistente en cada uno de los idiomas. Por ejemplo, configurar la alarma o saber el tiempo que va a hacer. Cada comando relevante contará como una funcionalidad.
- **Capacidad de escribir texto**: Capacidad de traducir a texto la conversación dictada.
- **Otras funcionalidades (+n)**: Otras funcionalidades no generales como pueden ser reconocimiento de entidades, stemming, autocorrección, etc.

Traducción automática

- **Corrección de gramática**: Si el traductor tiene autocorrector gramatical.
- **Idiomas desde los que puede traducir (+n)**: Cuando se está traduciendo al inglés o al español, el número de idiomas desde el que se puede traducir. Cada idioma contará como una funcionalidad.
- **Detección de idioma**: Cuando el idioma de partida es inglés o español, si es capaz de detectar el mismo.
- **Posibilidad de modificar traducciones (+1)**: Si permite modificar el texto traducido a posteriori para uso personal o si permite modificar el texto como sugerencia para el traductor. Cada una cuenta como una funcionalidad.
- **Versiones de la traducción**: Si el traductor sugiere distintas versiones para la traducción.
- **Traducción de archivos**: Capacidad de traducir archivos completos sin alterar los formatos de los mismos.
- **Traducción web**: Capacidad de traducir páginas web en su totalidad sin alterar el diseño de las mismas.
- **De texto a texto**: Traducción de texto a texto.
- **De texto a voz**: Traducción de texto a voz.
- **De voz a texto (+1)**: Traducción de voz a texto. Cuenta como funcionalidad adicional si puede hacerlo en tiempo real.
- **De voz a voz**: Traducción de voz a voz.
- **De imágenes de palabras a texto (+1)**: Capacidad de detectar texto en imágenes y traducirlo. Cuenta como funcionalidad adicional si puede hacerlo en tiempo real.

- **Adaptación a dominios**: Si el traductor puede adaptarse de manera automática a distintos dominios.
- **Personalización a dominios**: Posibilidad de añadir términos y datos de entrenamiento para adaptar el traductor a dominio específico.
- **Variantes regionales (+n)**: Si ofrece traducciones distintas para variantes regionales. Contará como una funcionalidad por variante.

Teclados predictivos

- **Corrección de gramática**: Si el teclado predictivo tiene autocorrector gramatical.
- **Predicción personalizada**: Si la predicción del teclado predictivo se ajusta a la manera de escribir del usuario.
- **Detección de idioma**: Cuando el idioma de partida es inglés o español, si es capaz de detectar el mismo.
- **Autocompletado de palabras**: Capacidad de autocompletar la palabra iniciada.
- **De voz a texto**: Escritura de voz a texto.
- **Sugerencias de palabras**: Capacidad de sugerir la próxima palabra a escribir.
- **Generación de texto (+2)**: Si puede sugerir, no solo palabras, sino textos completos o snippets. Se definen las siguientes funcionalidades adicionales: si puede sugerir expresiones, si puede sugerir frases completas y si puede sugerir párrafos o textos más extensos.

Buscadores web

- **Corrección de gramática**: Si el buscador tiene autocorrector gramatical.
- **Detección de significado**: Si el buscador tiene procesos de PLN que buscan por significado de la frase y no por palabras.
- **Clasificación de tema (+n)**: Capacidad de clasificar los documentos por tema. Si la solución es capaz de realizar clasificaciones siguiendo distintos criterios (tema, finalidad, etc.), contarán como una funcionalidad más por cada criterio adicional.
- **Detección de entidades**: Si es capaz de realizar la detección de entidades (nombres, sitios, empresas etc.) en los documentos.
- **Búsqueda de sinónimos**: Capacidad de utilizar sinónimos además de las palabras que se han incluído en la búsqueda.
- **Búsqueda de imágenes**: Capacidad de buscar imágenes partiendo de un texto introducido en el buscador.
- **Búsqueda de texto en imágenes**: Capacidad de buscar texto en documentos en formato imágen o en imágenes con texto.
- **Búsqueda de vídeos**: Capacidad de buscar vídeos partiendo de un texto introducido en el buscador.
- **Búsqueda de audio**: Capacidad de buscar en archivos de audio partiendo de un texto introducido en el buscador.
- **Búsqueda de respuestas**: Capacidad de devolver directamente una respuesta en lugar de una lista de URLs.
- **Búsqueda de información**: Capacidad de devolver información estructurada en lugar de una lista de URLs, como por ejemplo los knowledge panel de Google.

4.2.3. Indicador de brecha en funcionalidades

Dentro del ámbito de soluciones de mercado, se empleará un único indicador *Brecha en funcionalidades* (I.S.1) que capturará la brecha en cuanto a cobertura de funcionalidades en cada idioma ofrecidas por los productos. Este ámbito será independiente de aspectos subjetivos como satisfacción de usuario o de medidas de efectividad. Para cada familia de aplicaciones, el indicador I.S.1 se calculará de la siguiente forma. Sea F_I^p y F_E^p el conjunto de funcionalidades presentes en el producto p desarrollado para el inglés o español respectivamente:

$$\text{I. S. 1}(p) = \frac{|F_I^p \setminus F_E^p| - |F_E^p \setminus F_I^p|}{|F_I^p \cup F_E^p|} \cdot 100$$

Esta definición cumple las siguientes propiedades. La brecha es nula si ambas lenguas ofrecen las mismas funcionalidades:

$$F_I^p = F_E^p \Longrightarrow \text{I. S. 1}(p) = 0$$

El indicador es simétrico respecto a los idiomas. Además, la brecha es del 100 % sólo si ninguna de las funcionalidades quedan cubierta en español.

$$F_E^p = \emptyset \Longleftrightarrow \text{I. S. 1}(p) = 100\,\%$$

Dada una diferencia fija de funcionalidades en ambos sentidos, la brecha es inversamente proporcional al número de funcionalidades presentes en alguna de los idiomas:

$$|F_I^p \setminus F_E^p| = k \wedge |F_E^p \setminus F_I^p| = k' \Longrightarrow \text{I. S. 1}(h) \propto \frac{1}{|F_I^p \cup F_E^p|}$$

En caso de que el producto cubra todas las funcionalidades en inglés, entonces el indicador será proporcional a la cantidad de funcionalidades no cubiertas en español:

$$F_I^p \supset F_E^p \Longrightarrow \text{I. S. 1}(p) \propto |F_I^p| - |F_E^p|$$

Indicador S.1: Brecha en funcionalidades

Representa la diferencia en cuanto a funcionalidades ofrecidas por los productos en ambos idiomas. Sea H el conjunto familias de aplicaciones consideradas en el proyecto, sea W_h el peso asignado a una de las familias de aplicaciones, y sea P_h el conjunto de productos identificados para dicha familia de herramientas:

$$\text{I. S. 1} = \sum_{h \in H} W_h \frac{1}{|P_h|} \sum_{p \in P_h} \frac{|F_I^p \setminus F_E^p| - |F_E^p \setminus F_I^p|}{|F_E^p \cup F_I^p|}$$

F_E^p y F_I^p representan el conjunto de funcionalidades presentes en el producto p en inglés y español respectivamente.

4.3. Indicadores Ámbito 3: Nivel de adopción

En los últimos tiempos, la Inteligencia Artificial (IA) basada en lenguaje se ha convertido en un tema cada vez más relevante en el mundo, y se espera que su impacto en la forma en que las empresas operan y compiten continúe creciendo en el futuro. En este trabajo nos centramos en dar respuesta a dos preguntas fundamentales: ¿Se están adoptando estas herramientas de manera dispar en las empresas de habla hispana e inglesa? ¿Existen diferencias de adopción entre los ciudadanos de ambos idiomas?

Para contestar estas preguntas, se realiza un análisis de la adopción de herramientas de tecnologías del lenguaje en empresas de habla hispana e inglesa, utilizando una variedad de fuentes y métodos. En primer lugar, se identifican una serie de empresas representativas y se analizan menciones de herramientas de IA en sus presentaciones e informes de resultados corporativos. A continuación, se realiza un análisis similar de las menciones en medios de comunicación y se calcula el impacto de la adopción de tecnologías del lenguaje. Además, se llevan a cabo encuestas en España y Estados Unidos para conocer el grado de adopción de herramientas por parte de los ciudadanos.

Para medir el nivel de adopción de las soluciones (las cuales fueron indicadas en el documento “Ambito 2 Soluciones de Mercado Informe Año 2”) y las tecnologías por parte de las empresas se han utilizado

las empresas indicadas en la Tabla 1 como referencia. Se han seleccionado 20 compañías del IBEX-35 (Insider Inc., 2023a) y 20 compañías del S&P 500 (Insider Inc., 2023b). El criterio de selección ha sido el de mayor capitalización bursátil por medio de índices conocidos. De esta manera se capta la realidad de las grandes empresas en cada uno de los países sin caer en una selección subjetiva.

Tabla 1: Empresas con las que se medirá el nivel de adopción (indicadores A.1, A.2, A.3, A.4 e A.5)

	IBEX 35	S&P 500
1	Inditex	Apple
2	Iberdrola	Microsoft
3	Santander	Alphabet
4	BBVA	Amazon
5	Caixabank	NVIDIA
6	Naturgy	Meta
7	Amadeus	Berkshire Hathaway
8	Telefonica	Tesla
9	Aena	Lilly
10	Cellnex	UnitedHealth Group
11	Endesa	Visa
12	Repsol	Walmart
13	ArcelorMittal	Exxon Mobil
14	ACS	JPMorgan Chase
15	Red Eléctrica	Johnson & Johnson
16	IAG	Procter & Gamble
17	Grifols	Broadcom
18	Acciona	MasterCard
19	Mapfre	Oracle
20	Sabadell	Home Depot

En cuanto a las tecnologías de la lengua, para poderlas identificar en los informes y medios, se ha utilizado el listado de palabras y expresiones que las representa, indicado en la Tabla 25 del Apéndice A. Al confeccionar la lista se ha prestado especial atención en evitar sesgos entre un idioma y el otro.

4.3.1. Indicadores de menciones en informes corporativos y medios

Las presentaciones de los resultados corporativos analizadas pertenecen a los tres años fiscales previos al año del informe. En el cálculo del indicador del año n, se utiliza los informes de los años n-2, n-3 y n-4 anteriores, ya que serán los tres últimos años en los que hay resultados publicados. Si alguna empresa tuviera un año fiscal que no terminara el 31 de diciembre, se considerará como del año n-2 las cuentas que se cierren en una fecha más próxima al 31 de diciembre de dicho año. Las menciones en medios son las detectadas en los tres años naturales anteriores (n-1, n-2 y n-3).

Teniendo en cuenta la distinta legislación entre países a la hora de publicar las cuentas, la documentación a estudiar será la siguiente:

- En el caso de España, se utilizaran los informes integrados y las memorias consolidadas de cuentas anuales, o documentos equivalentes publicados por la empresa.
- En el caso de USA, se utilizan las memorias anuales corporativas (cuando estén disponibles), el formulario 10-K y las declaraciones de poder del año siguiente.

En este ámbito, mediante los indicadores A.1, A.2, A.3 y A.4, se estudian las menciones de productos y tecnologías del lenguaje. Para la identificación de menciones en informes se extrae el texto utilizando una

solución de extracción de snippets. Para contabilizar el número de menciones en medios de comunicación se utilizará la herramienta Brandwatch. En ambos casos se trabajará con una bolsa de palabras.

Consideraremos como menciones de tecnologías de PLN los desarrollos propios de las empresas y adquisiciones o inversiones realizadas por las empresas, que podrán ser para el uso en procesos internos o para incluirlo en los productos o servicios ofertados. Para la identificación de menciones de estas tecnologías en informes y medios, se parte de un listado de palabras y expresiones que identifiquen tecnologías de la lengua. Se ha prestado especial atención en evitar sesgos entre el inglés y el español a la hora de seleccionar esta lista. Los grupos de tecnologías seleccionados son los siguientes:

Grupos de tecnologías

- Análisis de sentimiento
- Análisis sintáctico
- Análisis de texto
- Análisis morfológico
- Comprensión del lenguaje natural
- Generación de resúmenes
- Lingüística de corpus
- Lingüística estadística
- Menciones de PLN no específicas
- Modelado de lenguaje
- Procesamiento de voz
- Reconocimiento de entidades
- Reconocimiento óptico de caracteres
- Respuesta a preguntas
- Semántica del discurso
- Semántica léxica
- Semántica relacional
- Sistemas de diálogo
- Sistemas generativos
- Traducción

Todos los términos empleados se encuentran en el apéndice A.

Sea M_I y M_E el número de menciones encontradas en los respectivos idiomas, los indicadores de mención en medios e informes se computarán como:

$$\mathrm{Ind}(I,E) = \frac{M_I - M_E}{max(M_I, M_E)}$$

Este indicador tiene las siguientes propiedades. En primer lugar, es simétrico respecto a las lenguas ($\mathrm{Ind}(I,E) = -\mathrm{Ind}(E,I)$). Además, el indicador será cero si aparecen el mismo número de menciones en ambas lenguas.

$$M_I = M_E \Longrightarrow \mathrm{Ind}(I,E) = 0\,\%$$

El indicador será del 100 % si en ambas lenguas las menciones en inglés duplican a las menciones en español.

$$M_I = 2 \cdot M_E \Longrightarrow \mathrm{Ind}(I,E) = 100\,\%$$

En caso de una diferencia fija entre menciones y entre menciones para el inglés, el indicador será inversamente proporcional al número total de menciones en inglés.

$$M_I - M_E = k > 0 \Longrightarrow \mathrm{Ind}(I,E) \propto \frac{1}{M_I}$$

Definimos distintos indicadores para informes y medios respectivamente. Dada la baja frecuencia de menciones en informes, no se pondera por productos, familias de aplicaciones o dominios.

Indicador A.1: Brecha en menciones de productos en informes

Sea Inf_I y Inf_E la cantidad de menciones de cualquiera de los productos y familias de aplicaciones consideradas en el proyecto en informes:

$$\text{I. A. 1} = \frac{Inf_I - Inf_E}{max(Inf_I, Inf_E)} \cdot 100$$

Indicador A.2: Brecha en menciones de tecnologías de la lengua en informes

Sea Inf_I y Inf_E la cantidad de menciones de cualquiera de los productos y familias de aplicaciones consideradas en el proyecto en informes:

$$\text{I. A. 2} = \frac{Inf_I - Inf_E}{max(Inf_I, Inf_E)} \cdot 100$$

Por otro lado, las menciones de productos en medios, por la disponibilidad de datos, se ponderan en base a la familia de aplicaciones a la que pertenece el producto, dando el mismo peso a todos los productos seleccionados para cada familia de aplicaciones. Para evitar sesgar las medidas por la popularidad de las empresas o la capacidad de generar noticias de los medios en un país u otro, las menciones se normalizan dividiendo por el número total de noticias de las empresas de cada índice.

Sea med_I^i y med_E^i el número de menciones totales encontradas del producto o tecnología i en los respectivos idiomas los indicadores de mención en medios, y med_I y med_E el número de menciones totales encontradas (todas las noticias que mencionan a las empresas representativas de cada idioma). Las menciones normalizadas Med_I^i y Med_E^i se computan como:

$$Med_I^i = \frac{med_I^i}{med_I}$$

$$Med_E^i = \frac{med_E^i}{med_E}$$

Indicador A.3: Brecha en menciones de productos en medios.

Sea H el conjunto familias de aplicaciones consideradas en el proyecto, sea W_h el peso asignado a una de las familias de aplicaciones, y sea P_h el conjunto de productos identificados para dicha familia de aplicaciones:

$$\text{I. A. 3} = \sum_{h \in H} W_h \frac{1}{|P_h|} \sum_{p \in P_h} \frac{Med_I^p - Med_E^p}{max(Med_E^p, Med_I^p)}$$

Med_E^p y Med_I^p representan el número de menciones del producto p encontradas en medios.

Indicador A.4: Brecha en menciones de tecnologías en medios.

Sea H el conjunto familias de aplicaciones consideradas en el proyecto, sea W_h el peso asignado a una de las familias de aplicaciones, y sea T_h el conjunto de menciones identificadas para dicha familia de aplicaciones:

$$\text{I. A. 4} = \sum_{h \in H} W_h \frac{1}{|A_h|} \sum_{t \in T_h} \frac{Med_I^t - Med_E^t}{max(Med_E^t, Med_I^t)}$$

Med_E^t y Med_I^t representan el número de menciones la tecnología t encontradas en medios.

Se considera que la adopción de un grupo de tecnologías tiene impacto en una empresa si al menos existe una misma noticia que menciona dicha tecnología y la empresa. El impacto que tiene la adopción de cada uno de los grupos de tecnologías se ha calculado aplicando el criterio experto de científicos de datos y atendiendo a las siguientes características:

1. **Proximidad al mercado:** se refiere a lo cerca que se encuentra una tecnología de ser adoptada en el mercado. Esta proximidad puede ser alta si ya existen soluciones similares en el mercado o si no requiere de otras tecnologías para llegar a construir una solución; y será baja si no existen soluciones comercialmente viables en el mercado o si requiere la combinación de muchas otras tecnologías para llegar a construir una solución.

2. **Viabilidad técnica:** se refiere a si la tecnología puede ser desarrollada y utilizada en la práctica, teniendo en cuenta factores como la disponibilidad de recursos técnicos y financieros, la complejidad del desarrollo y los posibles obstáculos técnicos que puedan surgir durante el proceso de implementación.

3. **Demanda de mercado:** se refiere a la cantidad de clientes potenciales que están interesados en la tecnología y a la cantidad de ventas o ingresos que se pueden esperar de la adopción de la tecnología.

4. **Potencial disruptivo:** se refiere a la capacidad de la tecnología para alterar significativamente un mercado o industria existente, ya sea creando nuevas oportunidades o amenazando la existencia de soluciones y productos existentes.

Siguiendo este criterio, se ha otorgado la categoría mostrada en la Tabla 2 a cada una de las tecnologías identificadas.

Tabla 2: Impacto asignado a cada una de las tecnologías.

Grupo de tecnologías	**Impacto**
Análisis de sentimientos	alto
Análisis sintáctico	bajo
Análisis de texto	bajo
Análisis morfológico	bajo
Comprensión del lenguaje natural	muy alto
Generación de resúmenes	medio
Lingüística de corpus	muy bajo
Lingüística estadística	bajo
Menciones de PLN no específicas	medio
Modelado de lenguaje	bajo
Procesamiento de voz	alto
Reconocimiento de entidades	medio
Reconocimiento óptico de caracteres	medio
Respuesta a preguntas	alto
Semántica del discurso	medio
Semántica léxica	bajo
Semántica relacional	bajo
Sistemas de diálogo	alto
Sistemas generativos	alto
Traducción	alto

Para otorgar un valor numérico a cada una de las categorías se ha seguido el criterio de medición de impacto del modelo de puntuación RICE.[7] El modelo de puntuación RICE es un marco diseñado para determinar qué productos, características y otras iniciativas priorizar según cuatro factores: alcance, impacto, confianza y esfuerzo. Siguiendo el criterio establecido por el modelo para la medición del impacto, se dará el siguiente **valor numérico** a cada una de las categorías: Muy bajo: 0.25; Bajo: 0.5; Medio: 1; Alto: 2; Muy alto: 3.

Se aplica un indicador análogo a los indicadores A.1-4 y heredando por tanto sus propiedades.

Indicador A.5: Brecha en impacto de las tecnologías en la empresa.

Sea M_I y M_E el conjunto de empresas establecidas en territorio de habla inglesa y española respectivamente. sea Imp_m el impacto producido por el uso de tecnologías de la lengua en la empresa m:

$$\text{I. A. 5} = \frac{Imp_I - Imp_E}{max(Imp_I, Imp_E)}$$

donde

$$Imp_I = \frac{1}{|M_I|} \sum_{m \in M_I} Imp_m \qquad Imp_E = \frac{1}{|M_E|} \sum_{m \in M_E} Imp_m$$

[7] https://www.productplan.com/glossary/rice-scoring-model/

4.3.2. Indicadores de encuestas de adopción

El indicador A.6 refleja el nivel de adopción para uso profesional de soluciones en inglés respecto al español. Se considera la adopción para uso profesional de una solución si un encuestado manifiesta haber utilizado la solución para uso profesional o para uso profesional y personal. Este indicador se obtendrá por medio de encuestas,

Con el propósito de medir la experiencia de usuario y el nivel de adopción y, de esta manera, complementar el análisis de redes sociales, notas de prensa e informes corporativos, se han realizado 901 encuestas en España y 904 en Estados Unidos sobre las soluciones seleccionadas de cada una de las áreas. Los encuestados han sido preguntados por todas la áreas. Por lo tanto, el número de encuestas por área e idioma es de 901. Esta cifra se ha establecido con el propósito de conseguir resultados estadísticamente significativos a nivel global.

En las encuestas se han incluido preguntas relativas a las siguientes áreas de soluciones:

- Análisis de opiniones
- Asistentes virtuales
- Traducción automática
- Teclados predictivos
- Buscadores web

Además, se ha recogido información sociodemográfica del encuestado con el objetivo de poder valorar la representatividad de la muestra. En cuanto a la adopción de las soluciones, se discrimina entre el uso personal (adopción ciudadana), el uso profesional (adopción empresarial) y el uso combinado (personal y profesional). El listado de preguntas que conforman las encuestas realizadas primero en España y luego en Estados Unidos se adjuntan en el Apéndice C.

Se aplicará un indicador análogo a los anteriores, heredando sus propiedades.

Indicador A.6: Brecha en adopción para uso profesional.

Sea H el conjunto familias de aplicaciones consideradas en el proyecto, sea W_h el peso asignado a una de las familias de aplicaciones, y sea P_h el conjunto de productos identificados para dicha familia de aplicaciones:

$$\text{I. A. 6} = \sum_{h \in H} W_h \frac{1}{|P_h|} \sum_{p \in P_h} \frac{GAE_I^p - GAE_E^p}{max(GAE_E^p, GAE_I^p)}$$

donde GAE_E^p y GAE_I^p representan la adopción para uso profesional del del producto p en ambas lenguas.

De manera análoga, el indicador A.7 identifica mediante cuestionarios la brecha de adopción ciudadana, es decir, productos de uso personal.Se considerará la adopción para uso personal una solución si un encuestado manifiesta haber utilizado la solución para uso personal o para uso profesional y personal. El diseño de las encuestas para este indicador se describe detalladamente en el apéndice C.

Indicador A.7: Brecha en adopción para uso personal.

Sea H el conjunto familias de aplicaciones consideradas en el proyecto, sea W_h el peso asignado a una de las familias de aplicaciones, y sea P_h el conjunto de productos identificados para dicha familia de aplicaciones:

$$\text{I. A. 6} = \sum_{h \in H} W_h \frac{1}{|P_h|} \sum_{p \in P_h} \frac{GAC_I^p - GAC_E^p}{max(GAC_E^p, GAC_I^p)}$$

donde GAC_E^p y GAC_I^p representan la adopción para uso personal del del producto p en ambas lenguas.

4.4. Indicadores Ámbito 4: Experiencia de usuario

En este ámbito, la experiencia de usuario se mide por medio de cuatro indicadores, dos de los cuales son calculados a partir de las opiniones y reseñas que los usuarios dejan de las soluciones que éstos utilizan y

otros dos indicadores que son calculados a partir de las respuestas que se obtienen de cuestionarios que se realizan a usuarios de las soluciones en cuestión.

4.4.1. Indicadores de análisis de opiniones

Para el primer indicador (E.1) denominado *brecha en polaridad reputacional*, se analizan entradas, opiniones o reseñas escritas en español e inglés para productos en español y en inglés respectivamente. Las fuentes que se han seleccionado para obtener las opiniones del posterior análisis son de tres tipos: redes sociales, comunidades web y reseñas. A continuación se listan las fuentes que se han utilizado para la extracción automática de opiniones y comentarios de las soluciones analizadas.

Para la obtención de los **mensajes de redes sociales** se ha utilizado la herramienta Brandwatch[8] que permite la extracción de datos mediante consultas. En las herramientas donde el uso principal se corresponde con la funcionalidad que se quiere estudiar, por ejemplo, el caso de la funcionalidad de traducción de DeepL, la consulta se ha realizado con el propósito de obtener todas las opiniones sobre dicha marca. En las herramientas cuyo uso principal no es el de la funcionalidad que se quiere estudiar, por ejemplo, el caso de la funcionalidad de escritura predictiva de GMail, Outlook, Google Workspace o Microsoft Word, la consulta se ha realizado con el propósito de obtener los mensajes que mencionan tanto la marca como la funcionalidad, sacrificando cobertura para tener mayor precisión.En todos los casos en que el nombre de la solución pudo resultar ambiguo, como lo son el de Alexa o Resonate, se han tomado medidas para desambiguar su nombre acompañándolo por otros términos o expresiones utilizadas para hacer referencia a la solución. En el caso de Alexa lo son por ejemplo el término “amazon” o “asistente virtual” (“virtual assistant” en inglés).

Para la obtención de las reseñas, en primer lugar, se han identificado aquellas soluciones que tienen aplicaciones móviles en Google Play[9] o App Store.[10] En segundo lugar, se ha buscado si la solución tiene una página de reseñas en G2.[11] Por último, se han utilizado *scrapers* desarrollados por el área de Deep Learning[12] de LLYC[13] para extraer las reseñas de todas las aplicaciones móviles identificadas y un *scraper* de Apify[14] para las páginas de reseñas de G2 encontradas.

En total se han analizado 576.463 opiniones del año 2023 de las fuentes seleccionadas. El detalle para cada uno de los idiomas y fuentes puede encontrarse en la Tabla 22.

Tabla 3: Número de mensajes y reseñas analizados por idioma.

Fuente	Español	Inglés
Twitter	49.470	109.597
Reddit	3.715	56.567
Tumblr	1.651	39.621
Blogs	16.588	42.350
Foros	9.942	56.016
Google Play	74064	82.249
App Store	2.201	28.370
G2	406	3.656
Total	**158.037**	**418.426**

Para el análisis se han seleccionado aquellas soluciones de las que se han logrado obtener, al menos, 100 opiniones en cada idioma. La lista final de las aplicaciones seleccionadas abarca herramientas de todas las áreas de aplicación y es la siguiente:

[8] https://www.brandwatch.com/
[9] https://play.google.com/store/apps
[10] https://www.apple.com/app-store/
[11] https://www.g2.com/
[12] https://llyc.global/en/capability/deep-learning/
[13] https://llyc.global/
[14] https://apify.com/

- Análisis de opiniones: Brandwatch, Digimind, Meltwater, NetBase Quid, Sprinklr, Talkwalker.
- Asistentes virtuales: Alexa, Bixby, ChatGPT, Google Assistant, Google Bard, Siri.
- Traducción automática: Microsoft Translator o Bing Translator, DeepL, Google Translate, memoQ Translator PRO, Smartling, Reverso Translation.
- Teclados predictivos: Fleksy, GBoard, GMail, iPhone Keyboard, Microsoft Office 365, Microsoft Swiftkey, Microsoft Outlook, Grammarly.
- Buscadores web: Bing, Brave Search, DuckDuckGo, Elasticsearch, Google Search, Perplexity, Yahoo Search.

El indicador de la brecha en polaridad reputacional sigue el mismo esquema que indicadores anteriores, considerando el ratio de la diferencia entre lenguas respecto al máximo entre ambas.

$$\frac{\mathrm{Pol}_I^h - \mathrm{Pol}_E^h}{max(\mathrm{Pol}_I^h, \mathrm{Pol}_E^h)}$$

La polaridad para cada idioma se computa como el ratio de opiniones positivas frente al total, considerando las muestras de opiniones neutras como elementos de incertidumbre con un peso de 1/2 para cada la polaridad negativa y positiva.

$$\mathrm{Pol}_E^d = \frac{\mathrm{Pol}_E^+ + \frac{1}{2}\,\mathrm{Pol}_E^N}{\mathrm{Pol}_E^+ + \mathrm{Pol}_E^- + \mathrm{Pol}_E^N}$$

Solo se tienen en cuenta aquellas aplicaciones que tienen un volúmen significativo de opiniones.

Indicador E.1: Brecha en polaridad reputacional

Sea H el conjunto familias de aplicaciones consideradas en el proyecto, sea W_h el peso asignado a cada una de ellas y sea P_h el conjunto de productos identificados para dicha familia de aplicaciones:

$$\text{I. E. 1} = \sum_{h \in H} W_h \frac{1}{|P_h|} \sum_{p \in P_h} \frac{\mathrm{Pol}_I^p - \mathrm{Pol}_E^p}{max(\mathrm{Pol}_I^p, \mathrm{Pol}_E^p)},$$

siendo:

$$\mathrm{Pol}_E^p = \frac{\mathrm{Pol}_E^+ + \frac{1}{2}\,\mathrm{Pol}_E^N}{\mathrm{Pol}_E^+ + \mathrm{Pol}_E^- + \mathrm{Pol}_E^N} \qquad \mathrm{Pol}_I^p = \frac{\mathrm{Pol}_I^+ + \frac{1}{2}\,\mathrm{Pol}_I^N}{\mathrm{Pol}_I^+ + \mathrm{Pol}_I^- + \mathrm{Pol}_I^N}$$

donde Pol_E^+, Pol_E^-, Pol_E^N, Pol_I^+, Pol_I^- y Pol_I^N representan la cantidad de entradas positivas, negativas y neutras de productos en español o inglés respectivamente para el producto p.

El siguiente indicador (E.2) cuantifica la brecha en cuanto a atributos valorables del producto. Para ello, se considera la polaridad reputacional de menciones de productos relativas a cada uno de los atributos. Estas menciones se categorizarán en atributos del producto en base a la ocurrencia de ciertos términos clave que se han identificado mediante expresiones regulares que se encuentran en el apéndice B. Los atributos considerados en este estudio son:

Rendimiento: Se engloba en este atributo tanto el rendimiento en términos de calidad y precisión de la tarea a realizar como la velocidad de la solución. Podrá haber una brecha entre el inglés y el español si los modelos utilizados no tienen el mismo rendimiento.

Usabilidad: Aglutina todos aquellos términos que tienen que ver con la usabilidad de la solución: sencillez de uso, flexibilidad, posibilidad de personalización, compatibilidad, imagen visual etc. La brecha puede existir si las aplicaciones tienen distintas funcionalidades en cada uno de los idiomas.

Seguridad y privacidad: Se aglutinan todos los términos que tengan que ver con la seguridad y la privacidad de la solución. Por ejemplo, en Europa hay soluciones que tienen desactivadas algunas de las funcionalidades debido a la GDPR. Estas limitaciones, pueden generar una brecha negativa en atributos como el rendimiento y la usabilidad, y una brecha positiva en el de seguridad y privacidad.

Precio: Para aquellas soluciones de pago, o que tengan funcionalidades de pago, se valorará el atributo precio. Se incluirán valoraciones de lo caro o barato que resulta. La brecha en este atributo puede darse debido a diferentes precios por territorios o idiomas o por diferencias en el poder adquisitivo en los territorios de habla inglesa y habla hispana.

El indicador (E.2) sigue el mismo esquema que el indicador anterior. Las opiniones positivas o negativas para los atributos de todos los productos en un idioma y aplicación, se consideran indistintamente. Es decir, tendrán más peso en el indicador aquellos productos que sean más populares en las redes sociales. Se asignará el mismo peso a cada uno de los cuatro atributos.

Indicador E.2: Brecha en curvas de valor

Sea H el conjunto familias de aplicaciones consideradas en el proyecto, sea W_h el peso asignado a cada una de ellas, y sea C el conjunto de atributos considerador en las curvas de valor:

$$\text{I. E. } 2 = \sum_{h \in H} W_h \frac{1}{|C|} \sum_{c \in C} \frac{\text{Pol}_I^{h,c} - \text{Pol}_E^{h,c}}{max(\text{Pol}_I^{h,c}, \text{Pol}_E^{h,c})} ,$$

siendo:

$$\text{Pol}_E^{h,c} = \frac{\text{Pol}_E^{+} + \frac{1}{2}\,\text{Pol}_E^{N}}{\text{Pol}_E^{+} + \text{Pol}_E^{-} + \text{Pol}_E^{N}} \qquad \text{Pol}_I^{h,c} = \frac{\text{Pol}_I^{+} + \frac{1}{2}\,\text{Pol}_I^{N}}{\text{Pol}_I^{+} + \text{Pol}_I^{-} + \text{Pol}_I^{N}}$$

donde Pol_E^{+}, Pol_E^{-}, Pol_E^{N}, Pol_I^{+}, Pol_I^{-} y Pol_I^{N} representan la cantidad de entradas positivas, negativas y neutras de productos en español o inglés respectivamente para la familia de aplicaciones h y asociados al atributo c.

4.5. Indicadores de encuestas de experiencia de usuario

Con el propósito de medir la experiencia de usuario y el nivel de adopción y, de esta manera, complementar el análisis de redes sociales, notas de prensa e informes corporativos, se realizan encuestas en España y en Estados Unidos sobre las soluciones seleccionadas de cada una de las áreas.

Las encuestas miden el grado de adopción, el grado de satisfacción y las limitaciones que puedan tener las soluciones en el uso habitual. Se incluyen preguntas relativas a las cinco familias de aplicaciones escogidas (análisis de opiniones, asistentes virtuales, traductores automáticos, teclados predictivos y buscadores web). Además, se incluye información socio-demográfica del encuestado con el objetivo de poder valorar la representatividad de la muestra. Las preguntas específicas que se realizarán en España y en Estados Unidos se encuentran en el Apéndice C.

En las encuestas se aborda el grado de adopción de los productos, tanto en uso personal (adopción ciudadana) como profesional (adopción empresarial), el grado de satisfacción y las limitaciones de uso.

Indicador E.3: Brecha en grado de satisfacción

Sea H el conjunto familias de aplicaciones consideradas en el proyecto, sea W_h el peso asignado a cada una de ellas: Sea H el conjunto familias de aplicaciones consideradas en el proyecto, sea W_h el peso asignado a cada una de ellas y sea P_h el conjunto de productos identificados para dicha familia de aplicaciones:

$$\text{I. E. } 3 = \sum_{h \in H} W_h \frac{1}{|P_h|} \sum_{p \in P_h} \frac{\text{GS}_I^{p} - \text{GS}_E^{p}}{max(\text{GS}_I^{p}, \text{GS}_E^{p})} ,$$

donde GS_E^{p}, y GS_I^{p} representa el grado de satisfacción derivado de las encuestas para la familia de aplicaciones h en español e inglés respectivamente para el producto p.

Indicador E.4: Brecha en limitaciones

Sea H el conjunto familias de aplicaciones consideradas en el proyecto, sea W_h el peso asignado a cada una de ellas, y sea C el conjunto de atributos considerados en este indicador:

$$\text{I. E. } 4 = \sum_{h \in H} W_h \frac{1}{|P_h|} \sum_{p \in P_h} \frac{1}{|L|} \sum_{l \in L} \frac{\text{GL}_E^{p,l} - \text{GL}_I^{p,l}}{max(\text{GL}_I^{p,l}, \text{GL}_E^{p,l})} ,$$

donde $\text{GL}_E^{p,l}$ y $\text{GL}_I^{p,l}$, el ratio de usuarios que han observado la limitación l de las encuestas para el producto p en español e inglés respectivamente.

5. Cálculo de indicadores: Ámbito 1 - Estado del arte

En esta sección se presentan los resultados obtenidos dentro del Ámbito 1, Estado del Arte.

5.1. Cálculo de indicadores de diseminación

5.1.1. D.1: Brecha en publicaciones científicas [D.1: 98 %]

A continuación se detalla la metodología adoptada para obtener datos para el cálculo del indicador D.1 (brecha en publicaciones) correspondiente al segundo año.

- **Fuente de datos:** Se han considerado las publicaciones en los dos congresos de mayor prestigio internacional en las áreas de PLN (ACL, Association for Computational Linguistics)[15] y de la recuperación de información (SIGIR, Special Interest Group on Information Retrieval),[16] respectivamente. Ambas conferencias están indexadas en la base de datos SCIE, en la categoría 1, y en la base de datos CORE, en la categoría A*. Además, se han considerado las conferencias de índole internacional CONLL (Computational Natural Language Learning)[17], EACL (European Chapter of the Association for Computational Linguistics)[18], y NAACL (North American Chapter of the Association for Computational Linguistics). Además de las conferencias anteriores ya monitorizadas en el año anterior, se ha incorporado en esta iteración el congreso EMNLP, que el segundo congreso internacional de mayor importancia tras el ACL dentro del área de procesamiento de lenguaje natural. No hemos añadido en este indicador las publicaciones asociadas a las campañas de evaluación en donde se comparten datos de entrenamiento y test (CLEF, IBERLEF, SEMEVAL, etc.), pues estos datos se tendrán en cuenta en el indicador de brecha en datos anotados de entrenamiento.

- **Intervalo temporal:** En este informe, se ha considerado para el cálculo del indicador las publicaciones referidas a los últimos 5 años, de 2019 a 2023. Respecto al informe del año anterior, desplazamos un año la ventada temporal (2018-2022) con el fin de estudiar la evolución de los indicadores bajo las mismas condiciones.

- **Extracción de datos:** Se ha realizado una búsqueda semi automática de la palabra “Spanish” en el título y resumen (*abstract*) de los artículos publicados, para localizar aquellos que experimentaron sobre datos en español. En el caso de las campañas de evaluación se ha considerado el artículo principal publicado por los organizadores. Se han revisado manualmente los resultados obtenidos para validarlos y se ha comprobado si se trata de trabajos multi-lingües inglés-español. Para obtener las publicaciones que han experimentado sobre datos en inglés, se ha considerado como tales todos aquellos artículos que no mencionan el nombre de los idiomas más representados en la conferencia (“Spanish”, “French”, “German”,“Italian”, “Chinese”).

- **Fecha de la búsqueda**: 31 de diciembre de 2023.

- **Cálculo del indicador:** En este caso se ha calculado el Indicador D.1 sobre el total de publicaciones en cada congreso a lo largo de los años para cada uno de los idiomas.

La Tabla 4 muestra los resultados para congresos internacionales y las cifras en las que se ha basado el cómputo.

En base a estos datos, en el congreso SIGIR de recuperación de información hay una brecha del 100 %, en el congreso ACL de PLN una brecha del 99 % y en el congreso CoNLL centrado en lingüística computacional, una brecha del 99 %, siendo sensiblemente mayor que el calculado en el informe anterior. Al igual que en el informe del año anterior, la predominancia del inglés respecto a otros idiomas parece ser un fenómeno bastante general.

[15] https://aclanthology.org/venues/acl/
[16] https://sigir.org/
[17] https://www.conll.org/
[18] https://eacl.org/

La brecha entre el español y el inglés desciende sensiblemente en el caso de los congresos internacionales de índole europeo y americano como EACL o NAACL (97 % en ambos casos).

Con respecto al informe del año anterior, se se ha ampliado el espacio de búsqueda a nuevos congresos, en concreto los congresos EMNLP o NAACL, obteniéndose resultados similares a los del congreso homólogo ACL.

Como criterio de agregación consideramos la suma de publicaciones en todos los congresos (en los dos últimos años). La última parte de la tabla muestra los resultados agregados. **En global, el español presenta una brecha similar a otras lenguas europeas como el francés o el alemán, en torno al 98 %.** La brecha desciende en el caso del chino (93 %).

De nuevo, no se observa una evolución de reducción de brecha a lo largo de los años. De hecho, la brecha global aumenta sensiblemente, por lo que no parece que existe ninguna tendencia beneficiosa para el español en este sentido. Por otro lado, se planteó en el informe del año anterior desgranar los indicadores en foros internacionales y nacionales además de un análisis por áreas de aplicaciones. Sin embargo, la disponibilidad de muestras para el español no ha permitido realizar dicho desgranamiento de indicadores.

5.1.2. D.2: Brecha en proyectos subvencionados [D.2: 96 %]

Al igual que en el año anterior, para la elaboración del indicador de brecha en proyectos subvencionados, hemos accedido a dos fuentes de datos que cubren proyectos europeos y estadounidenses respectivamente. En el ámbito europeo, accedemos a CORDIS (Community Research and Development Information Service[19]), portal de la Comisión Europea que recoge los proyectos financiados por los programas marco de investigación e innovación de la Unión Europea. En el ámbito estadounidense accedemos al portal de NSF (National Science Foundation)[20], una agencia independiente federal fundada en 1950 para la promoción de la ciencia en E.E.U.U. La consulta se ha realizado el 31 de diciembre de 2023.

En ambos casos se han considerado proyectos iniciados entre el 1 de enero de 2019 hasta el 31 de diciembre de 2023. En el caso de CORDIS, se ha realizado la siguiente búsqueda "Natural Language Processing.ªnd "Spanish", para recuperar posibles proyectos sobre procesamiento del lenguaje en nuestro idioma, y la búsqueda "Natural Language Processing"para recuperar posibles proyectos sobre procesamiento del lenguaje en inglés. Al resultado de la búsqueda en inglés, se han restado aquellos que proyectos en los que aparece mencionado alguno de los siguiente idiomas: "German", "French", "Spanish.ªnd Çhinese". Como resultado, en CORDIS se han encontrado 9 proyectos con experimentación en español frente a 226 proyectos con experimentación en inglés. Esto supone una brecha del 93 % según el indicador D.2.

En el caso de NSF, la búsqueda se ha realizado considerando los projectos asignados al área textit-Division of Information & Intelligent Systems en el caso de NSF. Se han encontrado 0 proyectos con experimentación en español frente a 126 proyectos con experimentación en inglés. lo que supone una brecha del 100 %. **En promedio, podemos considerar una brecha del 96 % para el indicador D.2.**. De nuevo, respecto al año anterior no hay un decrecimiento de la brecha sino más bien un aumento sensible.

Siguiendo las sugerencias en el informe anterior se ha estudiado las temáticas de investigación en español frente a inglés. No se ha encontrado ningún patrón característico. Se ha identificado un único proyecto orientado excusivamente al español denominado *Making contracts more human with state-of-the-art natural language processing techniques in Spanish.*

[19] https://cordis.europa.eu/search/es
[20] https://www.nsf.gov/awardsearch/advancedSearch.jsp

Tabla 4: Número de publicaciones en conferencias internacionales, por lengua y año, que describen experimentación para las lenguas consideradas, desde 2019 a 2023.

	2019	2020	2021	2022	2023	Brecha
			ACL			
Spanish	4	3	0	8	8	99 %
French	3	2	0	2	11	99 %
German	16	5	6	5	30	98 %
Chinese	19	18	36	25	36	93 %
English	618	751	668	662	825	
			SIGIR			
Spanish	0	0	0	0	0	100 %
French	0	0	0	0	0	100 %
German	0	0	0	0	0	100 %
Chinese	0	10	2	6	0	98 %
English	224	330	380	334	165	
			CONLL			
Spanish	0	2	0	-	0	99 %
German	6	1	3	-	0	94 %
French	1	1	2	-	0	97 %
Chinese	3	0	3	-	1	97 %
English	51	88	50	45	-	74
			EACL			
Spanish	8	-	-	-	4	96 %
German	19	-	-	-	4	92 %
French	8	-	-	-	6	95 %
Chinese	7	-	-	-	3	96 %
English	285	-	-	-	266	
			NAACL			
Spanish	15	-	21	3	-	94 %
German	19	-	17	6	-	93 %
French	12	-	10	4	-	97 %
Chinese	30	-	16	13	-	90 %
English	348	-	414	419	-	
			EMNLP			
Spanish	9	8	7	2	9	98 %
German	24	19	21	3	10	96 %
French	10	11	14	2	10	98 %
Chinese	33	26	27	27	30	93 %
English	606	688	779	793	989	
			Aggregated			
Spanish	43	13	29	16	21	0.98 %
French	82	22	47	17	25	0.97 %
German	52	17	36	15	46	0.97 %
Chinese	112	54	96	84	70	0.93 %
English	2513	1857	2746	2670	2245	

5.2. Cálculo de indicadores de recursos

5.2.1. R.0: Indicador de texto disponible en internet [83 %]

La información que se ha usado para calcular este indicador en esta segunda iteración se muestra en la Tabla 5 y se corresponde con as fuentes empleadas en el año anterior. Se ha recabado información sobre las siguientes colecciones de textos: Wikipedia, Internet, Internet Archive, PubMed y CommonCrawl. La consulta de información se realizó el 19 de febrero de 2024. Algunas fuentes proporcionan información en números absolutos y otras en términos de porcentajes.

Tabla 5: Datos sobre volumen de texto disponibles en Internet en inglés y en español.

Colección	Medida	Búsqueda Esp.	Esp.	Búsqueda Ing.	Ing.	Fuente	R.0
Wikipedia	# artículos	https://es.wikipedia.org/wiki/Wikipedia_en_espa%C3%B1ol	1932427	https://es.wikipedia.org/wiki/Wikipedia_en_espa%C3%B1ol	6785774		56 %
Internet	% páginas		5.6		51.20	https://w3techs.com/technologies/history_overview/content_language	80 %
Internet Archive	# textos	https://archive.org/search?query=Language%3A%28spanish%29	49017	https://archive.org/search?query=Language%3A%28english%29	8234197	https://archive.org/	99 %
PubMed	# textos	https://pubmed.ncbi.nlm.nih.gov/?term=Spanish%5BLanguage%5D&sort=	385212	https://pubmed.ncbi.nlm.nih.gov/?term=English%5BLanguage%5D	32013562	https://pubmed.ncbi.nlm.nih.gov/	98 %
Common Crawl: CC-MAIN-2023-50	% páginas		4.5391		44.4285	https://commoncrawl.github.io/cc-crawl-statistics/plots/languages	81 %
promedio							**83 %**

Para el cálculo del indicador, se asigna el mismo peso a todas las fuentes. De este modo, **obtenemos una brecha promedio R.0 del 83 %**. Ésta es inferior a la obtenida el año anterior (84 %).

5.2.2. R.1: Indicador de modelos pre-entrenados [76 %]

En esta iteración se ha considerado de nuevo el repositorio Hugging Face[21], que contiene gran cantidad de modelos de lenguaje pre-entranados. La interfaz de búsqueda permite filtrar modelos pre-entrenados por lenguas y por categorías de tareas de PLN. La Tabla 6 muestra los resultados obtenidos, según la búsqueda por lenguas y por tareas, realizada el 19 de febrero de 2024. Los modelos bilingües (inglés/español) no están incluidos en el recuento de modelos por idioma individual. Por ello, las variables del indicador se han fijado de la siguiente forma: $|M_I|$ y $|M_E|$ se han calculado como la primera o la segunda columna más la tercera respectivamente para cada idioma. La variable $|M_I \bigcup M_E|$ se ha estimado como la suma de las dos primeras columnas mas la tercera.

Como muestran los resultados, teniendo en cuenta el total de modelos se obtiene una brecha del 84 %, frente a un 83 % en el año anterior. Si se promedia las brechas obtenidas por tarea de PLN, se obtiene un 76 %, frente a un 75 % obtenido el año anterior. Sin embargo, como muestra la figura 11, las diferencias a escala de tarea son muy variables, aumentando o disminuyendo la brecha dependiendo de la tarea. Al igual que el año anterior, la menor brecha se obtiene en el caso de la traducción automática, que siempre implica varias lenguas (38 %). Si descartamos la traducción automática, obtenemos un indicador promedio por tareas de 79 %. Es decir, asignamos el mismo peso P_d a todas las categorías de tareas en Hugging Face. **Por consistencia con el informe del año anterior, como valor final del indicador R.1 consideramos el promedio por tareas, obteniendo una brecha del 76 %.**

[21] https://huggingface.co/

Tabla 6: R.1: Resultados del indicador de modelos pre-entrenados basado en datos de Hugging Face.

	Inglés	Español	Ing./Esp.	R.1
Total	34142	2495	1227	84 %
Por categorías de PLN				
Generación de textos	13785	465	368	91 %
Clasificación de textos	3506	266	103	84 %
Generación de textos Text2Text	2087	190	83	80 %
Traducción automática	925	370	150	38 %
Enmascaramiento de palabras	730	108	53	70 %
Clasificación de tokens	983	216	52	61 %
Búsqueda de respuestas	569	44	8	85 %
Resumen automático	448	36	16	82 %
Similitud entre frases	356	56	40	66 %
Sistemas conversacionales	64	0	0	100 %
Clasificación Zero-shot	107	27	18	53 %
Búsqueda de respuestas sobre tablas	49	1	0	96 %
Promedio por tareas				**76 %**

5.2.3. R.2:Indicadores de datos anotados [R.2: 55 %; R.2.a: 81 %; R.2.b: 29 %]

En esta segunda iteración se han estimado los indicadores R.2.a y R.2.b siguiendo lo mismos criterios que en el primer año de proyecto. A continuación se detalla la metodología adoptada:

- **Selección de repositorios:** Se han considerado los repositorios LRE Map,[22] repositorio de recursos lingüísticos promovido y mantenido por la European Language Resources Association (ELRA), y el catálogo LDC,[23] Linguistic Data Consortium, organización con sede en la Universidad de Pennsilvania, que crea y distribuye recursos lingüísticos. Además, se ha considerado el repositorio Hugging Face, cuya interfaz de búsqueda permite filtrar conjuntos de datos anotados por tarea e idioma. Los tres repositorios son repositorios de referencia en el área del PLN y el aprendizaje automático.
- **Selección de campañas de evaluación:** Se han considerado los datasets generados en las campañas de evaluación más importantes en tareas de PLN, IberEVAL y SEMEVAL, que son de ámbito nacional e internacional respectivamente. Además se han considerado los datasets generados en los foros de evaluación de sistemas de acceso a la información CLEF y TREC, de ámbito europeo e internacional respectivamente.
- **Intervalo temporal:** Se han considerado para el cálculo del indicador los recursos liberados o publicados durante los últimos 5 años, de 2019 a 2023.
- **Fecha de búsqueda**: 31 de diciembre de 2023.
- **Extracción de datos:** Se ha realizado de forma manual, mediante búsquedas en los repositorios indicados y en los *proceedings* y *working notes* de las campañas de evaluación. Se han clasificado los distintos conjuntos de datos según su idioma (inglés y/o español) y dominio.

[22] https://lremap.elra.info/
[23] https://catalog.ldc.upenn.edu/

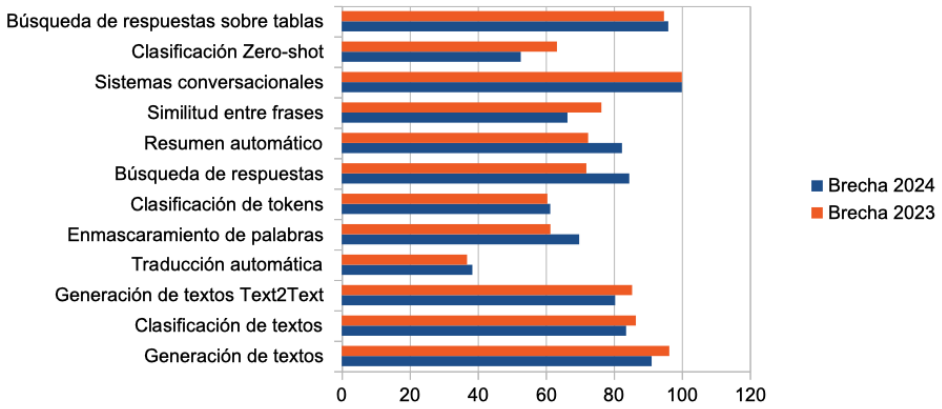

Figura 11: Comparativa de brechas en cantidad de modelos entrenados respecto al informe anterior.

En cuanto al indicador R.2.a (disponibilidad de corpus anotado para entrenamiento y test en foros internacionales), las búsquedas por idioma en el repositorio ELRA se han realizado sobre la categoría de recursos *"Dataset & evaluation data"* e introduciendo los filtros *Language="Spanish"* y *Language="English"* respectivamente. Sin considerar fecha, encontramos 213 recursos en español frente a 1773 recursos en inglés, lo que supone una brecha del 79 %. Asumimos que los recursos son de idioma único, por lo que el denominador del indicador es la suma de ambos números.

En cuanto al catálogo LCD, entre 2019 y 2023, encontramos 10, 19, 12, 7 y 7 datasets en inglés, frente a 3, 2, 4, 0 y 1 datasets en español. Esto supone una brecha promedio del 74 %. Si no se tiene en cuenta los años, encontramos 40 recursos en español frente a 410 en inglés. Esto supone una brecha del 82 %. Por coherencia con el repositorio ELRA y para evitar el efecto de la variabilidad entre años, tomamos este indicador.

En cuanto a Hugging Face, considerando el total de conjuntos de datos anotados, encontramos 5,879 en inglés frente a 555 en español, lo que supone una brecha del 83 %, que al igual que se mantiene cuando promediamos por tareas. Consideramos esté último indicador por el mismo motivo ya descrito en la sección anterior. Aplicamos el mismo peso P_r a todos los repositorios. **Calculando el promedio sobre los tres repositorios, obtenemos una brecha R.2.a del 81.17 %**. Esta brecha supera sustancialmente a la brecha obtenida en la iteración en el año anterior. Esto se debe a una explosión de recursos anotados en inglés. Como muestra la figura 12, la brecha en recursos anotados en Hugging face asciende considerablemente en todas las tareas.

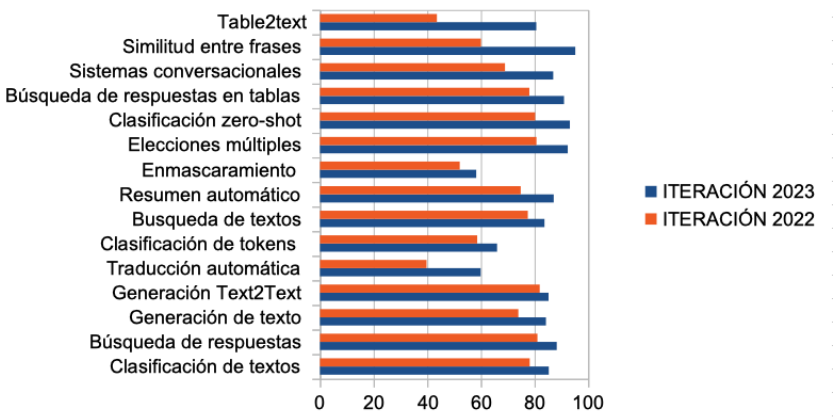

Figura 12: Comparativa de brechas en cantidad de recursos anotados en Hugging Face respecto al informe anterior.

Tabla 7: R.2.a: Indicador de conjuntos de datos anotados procedentes de repositorios.

	Inglés	Español	R.2.a
LRE			
Sin fecha	1773	213	**79 %**
LCD			
Sin fecha	410	40	**82 %**
2019	10	3	54 %
2020	19	2	81 %
2021	12	4	50 %
2022	7	0	100 %
2023	12	1	85 %
Promedio por años			74 %
Hugging Face			
Sin fecha	5,879	555	83 %
By NLP Task:			
Clasificación de textos	1310	104	85 %
Búsqueda de respuestas	802	50	88 %
Generación de texto	1157	99	84 %
Generación Text2Text	363	29	85 %
Traducción automática	343	86	60 %
Clasificación de tokens	259	53	66 %
Busqueda de textos	169	15	84 %
Resumen automático	480	33	87 %
Enmascaramiento	163	43	58 %
Elecciones múltiples	126	5	92 %
Clasificación zero-shot	141	5	93 %
Búsqueda de respuestas en tablas	127	6	91 %
Sistemas conversacionales	358	25	87 %
Similitud entre frases	81	2	95 %
Table2text	28	3	81 %
Promedio por tareas			**82 %**
Promedio por repositorios			**81 %**

En cuanto a recursos anotados en campañas de evaluación (indicador R.2.b), como muestra la Tabla 8, la brecha entre español e inglés en el caso de las campañas de evaluación en SEMEVAL, orientadas a PLN y de índole internacional, es del 78.95 %. En el caso de las campañas CLEF, orientadas a recuperación de información y de índole europeo, la brecha desciende hasta el 71.43 %. Al igual que en la iteración anterior, en ambos casos la brecha parece ser inferior o semjante a la de otros idiomas como el francés, alemán o chino. Por supuesto, en IBERLEF, conferencia que aglutina campañas de evaluación en el ámbito ibero-americano, la brecha se invierte, con 49 campañas sobre datos en español frente a 6 con datos en inglés. La brecha promedio para el español es del 24.06 %, promediando por campañas, es decir, asignando un p_c fijo para todas las campañas. Sin embargo, con el fin de tener en cuenta el tamaño de éstas (número de tareas), tomaremos como peso P_c de cada campaña su número de tareas, siendo esto equivalente a aplicar el indicador sobre la suma de las tareas en todas las campañas (internacional, europeo y nacional), **se obtiene una brecha del 28.57 % para el indicador R.2.b.**, ligeramente inferior a la brecha del 33 % obtenida en la iteración anterior. Esto se debe a una reducción sensible de la brecha en campañas

Tabla 8: Datos anotados en diferentes idiomas en las campañas de evaluación SEMEVAL, CLEF e IBERLEF.

	2019	2020	2021	2022	2023	R.2.c
			SEMEVAL-Workshops			
Español	1	1	0	1	3	78.95
Aleman	1	1	0	1	2	82.14
Francés	1	0	1	2	3	75.86
Chino	0	0	1	1	2	85.45
Inglés	8	10	9	12	12	
			CLEF			
Español	0	2	3	2	3	71.43
Francés	0	1	0	1	1	90.48
Aleman	1	1	1	2	0	84.62
Chino	0	0	0	0	0	100
Inglés	9	12	12	14	13	
			IBERLEF			
Español	8	7	11	9	14	-78.18
Inglés	0	0	1	1	4	
			Total			
Español	9	10	14	12	20	28.57
Inglés	17	22	22	27	29	
Brecha promedio inglés/español por campañas						24.06 %

de índole internacional (SEMEVAL) y europea (CLEF).

El promedio de R.2.a y R.2.b nos proporciona una **brecha en datos anotados R.2 del 55 %**, muy similar a la del año anterior (54 %), ya que el aumento de brecha en datos anotados en repositorios se compensa con la reducción de brecha en datos de campañas de evaluación.

5.3. Cálculo de indicadores de efectividad basados en experimentos [E1: 20±06 %]

5.3.1. Criterios de selección de datasets y tareas

- En cada dataset, los datos en inglés y en español se deben haber recolectado y anotado siguiendo la misma metodología, y el volumen de datos debe ser comparable entre ambos idiomas. No consideramos que cumplan este requisito los datasets que son traducción el uno del otro, ya sea automática o manual. En ambos casos pueden producirse sesgos en la evaluación (por supuesto, el supuesto de traducción automática introduce sesgos más acusados).

- En cada dataset, el subconjunto de test (sobre el que se evalúan los sistemas) no debe haber sido distribuido públicamente. De esta manera se evita el sobreajuste de datos y se minimiza la posibilidad de contaminación en los modelos (es decir, que el modelo haya visto las anotaciones manuales en la fase de preentrenamiento). La contaminación provoca una sobreestimación del rendimiento de un modelo contaminado con respecto a sus homólogos no contaminados. Desde un punto de vista científico las consecuencias son muy perjudiciales, ya que se publican conclusiones científicas erróneas.

- Para los datasets existentes, debe ser posible crear un subconjunto adicional de test privado utilizando la misma metodología con la que se construyó el dataset original. En este sentido, es preferible

adaptar datasets en los que podamos contar con el equipo original que los desarrolló y los anotó.

Otros aspectos que hemos considerado son los siguientes:

- La selección de tareas debe estar más orientada hacia las aplicaciones de la tecnología que hacia evaluaciones puramente lingüísticas. Las tareas más lingüísticas son adecuadas para evaluar el grado de conocimiento que tienen los modelos sobre la lengua, pero tienen una relación menos directa con el comportamiento de las aplicaciones de IA usadas por ciudadanos y organizaciones. Nuestro interés es realizar mediciones que tengan algún tipo de correspondencia con el uso práctico de estas tecnologías, así que queremos priorizar las tareas más cercanas al mercado de aplicaciones.

- Las tareas deben tener un grado de dificultad realista. Por un lado, hay datasets que pueden resolverse con un grado muy alto de acierto, pero no porque las tareas sean realmente sencillas, sino porque los sistemas aprenden los sesgos del dataset. En esos casos, el comportamiento de los sistemas fuera del laboratorio (es decir, sobre conjuntos de datos que no tienen los sesgos del conjunto de entrenamiento) es mucho peor que en las evaluaciones de laboratorio. En el otro extremo se encuentran los "diagnostic datasets", en los que se escogen ejemplos para el conjunto de test que requieran un conocimiento profundo del lenguaje para poder resolverse. Este tipo de datasets son muy útiles para identificar los puntos débiles de los modelos de lenguaje y para mejorarlos, pero son más difíciles que las situaciones promedio que nos encontramos en entornos de aplicación, de forma que tampoco ofrecen una imagen realista del rendimiento de las aplicaciones de Procesamiento del Lenguaje Natural. Nos interesa utilizar datasets en los que el rendimiento de los modelos de lenguaje se acerque al que encontraríamos en entornos reales de aplicación.

- Dificultad similar entre idiomas. Hemos establecido mecanismos para calibrar el leaderboard en función de la dificultad relativa intrínseca de los datasets en inglés y español, de manera que si para una tarea determinada hay una dificultad intrínseca (independiente del conocimiento del lenguaje que tengan los sistemas) diferente entre los dos idiomas, somos capaces de medirla y recalibrar la evaluación para que no contamine las diferencias observadas en el conjunto de test entre los dos idiomas. Sin embargo, conviene descartar los datasets en los que la dificultad intrínseca entre los dos idiomas es muy grande, porque esa es una señal de que la metodología con la que se han seleccionado y/o anotado en ambos idiomas seguramente no es equivalente.

- Los datasets y las tareas deben tener diversidad para ser representativos de las distintas aplicaciones del Procesamiento del Lenguaje Natural. Nos centraremos en tareas discriminativas que sean susceptibles de ser abordadas directamente por los modelos de lenguaje: en particular, de clasificación y de etiquetado. En el segundo año hemos empezado a trabajar también en tareas generativas por ser especialmente relevantes en el mercado actual de la IA, aunque su evaluación es mucho más compleja: o se invierte mucho esfuerzo en evaluaciones no automatizables, o se evalúan de forma poco precisa con medidas automatizadas de similitud textual con modelos realizados por humanos. También es deseable cierta diversidad en los dominios y en el tipo de textos.

- Accesibilidad. Los datos de entrenamiento deben ser fáciles de conseguir para los desarrolladores, idealmente mediante un sólo acuerdo centralizado con la UNED como proveedora.

Este conjunto de requisitos no se cumple en la mayoría de datasets disponibles, especialmente dada la necesidad que tenemos de expandirlos para disponer de un juego de pruebas privado que no haya sido publicado.

Los siguientes datasets se crearon para el leaderboard en el primer año de trabajo, y constituyen el leaderboard ODESIA CORE. En todos ellos se ha realizado al menos una anotación adicional para disponer de un juego de pruebas privado que no pueda ser leído por los modelos de lenguaje en su proceso de entrenamiento, garantizando así que la evaluación está libre de problemas de contaminación:

- DIPROMATS 2023. Este dataset fue creado desde cero para incorporarlo al Leaderboard ODESIA en la Versión 1. Se trata de un conjunto de tuits emitidos por diplomáticos de cuatro potencias mundiales (la Unión Europea, Rusia, China y Estados Unidos), anotados en función de las técnicas de propaganda que utilizan para transmitir una imagen determinada de sus países o de sus competidores a nivel global. Hay tres tareas asociadas con este dataset: identificación de propaganda, caracterización a grano grueso (cuatro técnicas) y caracterización a grano fino (15 técnicas subsumidas en las anteriores). Se trata de un problema de clasificación multiclase y multietiqueta. Se enmarca dentro de los problemas relacionados con la desinformación.

- EXIST 2022. Este dataset fue extendido para su incorporación al leaderboard en la Versión 1, creando un subconjunto adicional de datos anotados como conjunto de test privado. Se trata de un conjunto de tuits anotado en función de si contienen mensajes sexistas o no, y de qué tipo de sexismo se trata. Se enmarca dentro del problema de la toxicidad en redes sociales.

- DIANN 2023. Este dataset fue adaptado y extendido para su incorporación al Leaderboard. Se creó una partición de evaluación para la Versión 1 del Leaderboard. Los textos son resúmenes de artículos sobre biomedicina, y la tarea consiste en identificar menciones de discapacidades. Se trata por tanto de una tarea de etiquetación de secuencias.

En la Versión 2 del Leaderboard ODESIA CORE se han incorporado los siguientes datasets:

- EXIST 2023. Se trata de un dataset creado en su integridad para la Versión 2 del Leaderboard. Se compone de tuits etiquetados en función del tipo de sexismo expresado o descrito en ellos. Se trata, además, de un dataset desarrollado siguiendo el paradigma de "aprendizaje con desacuerdo" (Learning with Disagreement, LeWiDi) (Uma et al., 2021a), lo que lo convierte en el primer dataset para el entrenamiento y prueba de sistemas de detección de sexismo en textos construido conforme a este paradigma. Consta de tres particiones (entrenamiento, desarrollo, evaluación) y anotaciones para tres tareas: Detección de sexismo, categorización e identificación del emisor de sexismo. Se enmarca dentro del problema de la toxicidad en redes sociales.

- SQUAD/SQAC 2024. Este dataset contiene una partición de evaluación creada para la Versión 2 del Leaderboard ODESIA. Contiene artículos de divulgación científica del CSIC para el español y de Cambridge University para el inglés. La tarea que este dataset permite evaluar es la de comprensión de texto extractiva en sistemas de pregunta-respuesta. La tarea consiste en responder a preguntas sobre un texto, de tal manera que la respuesta sea un fragmento extraido directamente del texto. Se trata de una tarea de etiquetación de secuencias. El hecho de que los documentos anotados en los SQUAD/SQAC originales sean de fuentes distintas a los de nuestra anotación hace que, desde el punto de vista de los sistemas supervisados, este dataset sea particularmente complejo, ya que implícitamente se está midiendo la capacidad de transferencia del aprendizaje entre dominios. Además, es un dataset particularmente apropiado para evaluar modelos generativos en modo zero-shot (sin ejemplos) o few-shot (unos pocos ejemplos); de hecho, una de las motivaciones para incluirlo este año ha sido su proximidad a las aplicaciones más comunes a nivel empresarial de los modelos generativos: la capacidad de los modelos para extraer y sintetizar información de información corporativa en formato texto o semiestructurado, en una aproximación RAG (Retrieval-Augmented Generation) no supervisada.

La Tabla 9 muestra un resumen de los datasets del Leaderboard ODESIA CORE v2. Adicionalmente, se ha desarrollado otros datasets pensados exclusivamente para evaluar modelos generativos en modo zero-shot o few shot. De momento no se han añadido al leaderboard, ya que los modelos discriminativos no pueden abordarlos.

- UNED ACCESO 2024. El dataset contiene 1003 preguntas de opciones múltiples de once asignaturas del Curso de acceso para Mayores de 25 años de la UNED. Las preguntas y sus respuestas se han

Dataset	Tareas	Tarea abstracta	Dominio	Área de aplicación
DIANN 2023	detección de discapacidades	etiquetado	Biomedicina	Entidades nombradas
DIPROMATS 2023	Identificación de propaganda	Clasificación binaria	Geopolítica	Desinformación
	Caracterización de propaganda (gruesa)	Clasificación jerárquica multilabel	Geopolítica	Desinformación
	Caracterización de propaganda (fina)	Clasificación jerárquica multilabel	Geopolítica	Desinformación
EXIST 2022	Detección de sexismo	Clasificación binaria	Redes sociales	Toxicidad
	Categorización de sexismo	Clasificación jerárquica multiclase	Redes Sociales	Toxicidad
EXIST-2023	Identificación de sexismo	Clasificación binaria LeWiDi	Redes sociales	Toxicidad
	Intención de la fuente	Clasificación jerárquica LeWiDi	Redes Sociales	Toxicidad
	Categorización de sexismo	Clasificación jerárquica multilabel LeWiDi	Redes sociales	Toxicidad
SQUAD-SQAC 2024	Machine reading	Etiquetado de secuencias	Ciencia	Pregunta-respuesta

Tabla 9: Resumen de los datasets incorporados al Leaderboard ODESIA CORE v2.

traducido manualmente al inglés (sin intervención de ningún sistema de traducción automática, para evitar sesgos). Este dataset permite evaluar el conocimiento general de los modelos generativos, de forma similar a otros datasets como MMLU. Las diferencias con MMLU son: la disponibilidad de las preguntas en dos idiomas mediante traducciones manuales, y que el dataset no se hace público, de forma que se limitan los problemas de contaminación. Sobre este dataset se ha finalizado una experimentación exhaustiva sobre los mejores modelos generativos actuales (Claude 3 Opus y GPT-4) y sobre varios modelos abiertos (Llama-2, Mistral y Gemma).

- CURIA 2024. Este dataset contiene particiones de entrenamiento, desarrollo y evaluación. Los textos que lo conforman son sentencias de tribunales de justicia de la Unión Europea, que están acompañados de micro-resúmenes en lenguaje claro. En este caso, la tarea que permite evaluar el dataset CURIA-2024 es el resumen simplificado de textos legales, por tanto, se trata de una tarea de generación de texto. El dataset está finalizado y la experimentación se realizará en el tercer año de proyecto.

- Pron vs Prompt. El dataset consiste en 120 sinopsis para 60 títulos de películas imaginarias, 30 propuestos por GPT4 y 30 por un novelista de prestigio (Patricio Pron, premio Alfaguara de novela). Se solicitó a ambos, al escritor y GPT-4, que escribieran sinopsis de aproximadamente 600 palabras para cada título, incluyendo tanto los propuestos por ellos mismos como por su contraparte. En este caso, se está realizando una evaluación sistemática por parte de críticos y académicos, con la que se espera medir de forma precisa y fiable las capacidades de escritura creativa de los modelos de forma puntual.

En conjunto, los 5 datasets del Leaderboard ODESIA que tienen tests privados (ODESIA CORE) aportan 10 tareas distintas que abarcan problemas de clasificación (binaria, multiclase, multietiqueta, jerárquica, clasificación con disagreement) y etiquetado de secuencias. Hay varios tipos de textos: tuits en redes sociales, mensajes de autoridades y diplomáticos, resúmenes científicos y noticias de divulgación científica. Y en cuanto a dominios, se abarcan el biomédico, el de política y relaciones internacionales, las redes sociales, y dominios científicos variados.

Los detalles sobre la creación y contenido de cada dataset se especifican en los informes técnicos que también se entregan para cada dataset. A continuación se presenta un resumen de cada uno de ellos.

5.3.2. Datasets

Dataset 1: EXIST 2022

EXIST (sEXism Identification in Social neTworks) 2022 es un dataset desarrollado para facilitar la investigación en detección automática de sexismo en redes sociales. Se compone de textos cortos procedentes de redes sociales etiquetados en función del tipo de sexismo expresado o descrito en ellos. Contiene datos de dos redes sociales diferentes: Twitter [24] y Gab [25]. Se trata, por tanto, de mensajes cortos intercambiados en alguna de las dos redes anteriores y de dominio general (es decir, no versan, a priori, sobre ninguna temática en particular).

En total, el dataset se compone de 12,390 textos etiquetados: 6,226 en español y 6,164 en inglés. La distribución de los textos por partición (entrenamiento/test), fuente (Twitter/Gab) e idioma (inglés/español) se muestra en la Tabla 14. Sobre ellos se definen dos tareas:

- Tarea 1: detección de sexismo (clasificación binaria). Los sistemas deben decidir, para cada tweet, si contiene mensajes sexistas o no.
- Tarea 2: detección y caracterización de sexismo (clasificación multiclase). Los sistemas deben decidir, para cada tweet, si es o no sexista, y en caso afirmativo qué tipo de sexismo, entre las siguientes categorías: ideological and inequality, stereotyping and dominance, objectification, sexual violence, misogyny and non-sexual violence.

Entrenamiento				**Test**	
Twitter		**Gab**		**Twitter**	
Español	**Inglés**	**Español**	**Inglés**	**Español**	**Inglés**
5,211	5,152	490	492	522	513
10,363		982		1,035	

Tabla 10: Distribución de datos de EXIST 2022 por partición, fuente e idioma.

Cada texto del dataset tiene asignadas 1 o 2 etiquetas, dependiendo de si se trata de un texto sexista o no. La primera etiqueta responde a la pregunta: *¿Es el texto sexista, en cualquiera de sus formas, o describe conductas o situaciones en las que se produce discriminación sexista (es decir, es sexista en sí mismo, describe una situación sexista o critica un comportamiento sexista)?* En función de la respuesta a esta pregunta, la etiqueta puede tomar uno de los siguientes dos valores:

- `SEXIST`: el tuit es sexista en sí mismo, describe una situación sexista o critica un comportamiento sexista. Ejemplos de textos del dataset anotados con esta etiqueta son:
  (1) *Que materialistas se han vuelto las mujeres de hoy en día, aún recuerdo cuando las podíamos enamorar con puras mentiras.*
  (2) *Las mujeres no deberian ni maquillarse. Ya sabemos que son rompe bolas arregladas o desarregladas. Joda las amamos. JAJAJA.*
  (3) *Te tacharán de machista y misógino. Las mujeres pueden pegar los hombres no.*
- `NON-SEXIST`: el tuit no perjudica, discrimina o menosprecia a las mujeres ni se refiere a contenidos o situaciones en las que sí se haga. Ejemplos de textos del dataset anotados con esta etiqueta son:
  (4) *Mi mamá: Si te llegas a ir del país, vas a saber lo dura que es la vida, porque ni siquiera te dignas a fregar las cosas que usas.*
  (5) *Alguien por favor me regala un bate que yo entre los fachas los señoros de las nuevas masculinidades y los que mucho feminismo pero el porno no me lo toqueis quiero reventar una cabeza.*

[24] https://twitter.com/
[25] https://gab.com/

(6) *El impactante video de una mujer al volante que sale milagrosamente ilesa de un intento de asesinato...*

La segunda etiqueta responde a la pregunta: *De acuerdo con la faceta de la mujer que está siendo atacada, ¿qué tipo de sexismo encontramos en el tuit?*. El sexismo puede afectar a las mujeres en muchas facetas de sus vidas, incluidas las funciones domésticas y de crianza, las oportunidades profesionales, la imagen sexual y las expectativas de vida, por nombrar algunas. En función de la respuesta, la etiqueta puede tomar uno de los siguientes valores:

- `IDEOLOGICAL-INEQUALITY`: Descrédito ideológico, negación de la desigualdad y narrativa invertida. El tuit desacredita el movimiento feminista, rechaza la existencia de desigualdad entre hombres y mujeres, o presenta a los hombres como víctimas de la opresión de género. Ejemplos de tuits en esta categoría son:

  (7) *Mucho feminismo pero poco interés con casos de pedofilia, que hipócrita de mierda.*

  (8) *Mucho feminismo mucho feminismo pero andamos tirando historias burlándose de una piba solo porque el novio le metió los cuernos y volvió. Gente hipócrita en esta vida existen y muchas.*

  (9) *Mi hermana y mi madre se burlan de mí por defender todo el tiempo los derechos de todos y me acaban de decir feminazi, la completaron.*

- `STEREOTYPING-DOMINANCE`: Estereotipos y dominancia. El tuit expresa ideas falsas sobre las mujeres que sugieren que son más adecuadas o inapropiadas para ciertas tareas o afirma que los hombres son superiores a las mujeres. Ejemplos de textos del datasets con esta etiqueta son:

  (10) *Un taxista enojado con una mujer que pasaba en bicicleta le gritó "mujer tenias que ser" creo que la quiso insultar pero no se alto orgullo serlo bobo.*

  (11) *@RuloTanki @PrinncessRoyal Uy andas en tus días? Jajajja.*

  (12) *Paula2R faber_acuria A las mujeres hay que amarlas...solo eso... Nunca las entenderás.*

- `OBJECTIFICATION`: Objectificación. El tuit presenta a las mujeres como objetos al margen de su dignidad y personalidad, o asume o describe ciertas cualidades físicas que las mujeres deben tener para cumplir con los roles tradicionales de género. Ejemplos de textos anotados con esta etiqueta son:

  (13) *@EveryWhereTraps Tienes buen tamaño para ser chica, más tamaño ya no encuentras casi ropa y menos zapatos bonitos.*

  (14) *¿Cómo iba vestida?", "seguramente estaba provocando", "mmm ese short esta muy corto", "no te pongas eso que pareces una puta", "bueno iba sola por la calle a esa hora", "se lo busco", "bueno, para que toma entonces", "le gustó", "¿porque no gritó?*

  (15) *Pareces una puta con ese pantalón - Mi hermano de 13 cuando me vio con un pantalón de cuero.*

- `SEXUAL-VIOLENCE`: Violencia sexual. Se realizan sugerencias sexuales, solicitudes de favores sexuales o acoso de carácter sexual (violación o agresión sexual). Ejemplos de textos anotados con esta etiqueta son:

  (16) *@giulianarob @Gattarok @DemenArt ¿Misógino? Misoginia es odio a las mujeres y, en el porno los hombres las hacen disfrutar mucho.*

  (17) *#MeToo Estas 4 no han conseguido su objetivo. El juez estima que se abrieron de patas https://t.co/GSHiiwqY6Aánta lagartona hay en este https://t.co/8t5VmFIUFn.*

- `MISOGYNY-NON-SEXUAL-VIOLENCE`: Misoginia y violencia no sexual. El tuit expresa odio y violencia hacia la mujer. Ejemplos de textos anotados con esta etiqueta son:

(18) *Las mujeres de hoy en dia te enseñar a querer... estar soltero.*

(19) *Cualidades = odio a las mujeres, extranjeros, trabajadores, catalanes, vascos...*

(20) *Odio la misoginia más de lo que odio a las mujeres.*

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Informe del dataset EXIST 2022", que fue entregado en el Año 1.

Dataset 2: DIPROMATS 2023

DIPROMATS consiste en un conjunto de tweets en español y en inglés emitidos por diplomáticos y autoridades de cuatro potencias mundiales: EEUU, Rusia, China y la UE. Sobre ellos se ha realizado una anotación (de expertos) en la que se identifican los tweets que contienen algún tipo de propaganda, y una caracterización de las técnicas de propaganda usadas, tanto en grano grueso (cuatro categorías) como en grano fino (quince subcategorías). El dataset contiene 24,248 tweets anotados para las tres tareas, y se utiliza dentro del proyecto de dos formas: como parte del leaderboard bilingüe para evaluación comparada de modelos del lenguaje en español e inglés, y como parte de los datasets utilizados para la medición de la brecha de la IA entre ambos idiomas.

La Tabla 11 muestra la distribución de tuits y autoridades por áreas geopolíticas para el inglés, y la Tabla 12 para el español.

	China	Rusia	UE	EEUU	Total
Tuits	3,647	3,591	3,553	3,956	14,747
Autoridades	106	111	186	216	619

Tabla 11: Distribución del dataset DIPROMATS 2023 en inglés por áreas geopolíticas

	China	Rusia	UE	EEUU	Total
Tuits	2,997	1,391	2,465	2,738	9,501
Autoridades	25	22	48	40	135

Tabla 12: Distribución del dataset DIPROMATS 2023 en español por áreas geopolíticas

Utilizando las anotaciones manuales, se han definido tres tareas:

- **Tarea 1: Identificación de propaganda**, que se plantea como un problema de clasificación binaria que consiste en determinar si un tuit contiene o no técnicas de propaganda.
- **Tarea 2: Caracterización de la propaganda (grano grueso)**, que tiene como objetivo categorizar los mensajes propagandísticos según el tipo de propaganda que contienen. La categorización propuesta considera múltiples técnicas identificadas mediante revisión de la literatura existente, que se agrupan según sus características retóricas. Proponemos una tarea de clasificación multiclase y multietiqueta, en la que los sistemas tienen que asignar los tuits a una o más de las categorías definidas. Hay dos tipos de categorías:
  - Una categorización de **grano grueso** con cuatro clases de propaganda (más una clase negativa):
    - Grupo 0. `Not propaganda`
    - Group 1. `Appeal to Commonality`
    - Grupo 2: `Discrediting the opponent`
    - Group 3: `Loaded Language`
    - Grupo 4: `Appeal to authority`
- **Tarea 3: Caracterización de la propaganda (grano fino)**, idéntica a la anterior pero con un conjunto de subcategorías que refinan la clasificación anterior. En este caso se distinguen 15

subtipos de propaganda: `Flag Waving`, `Ad Populum / Ad antiquitatem`, `Name Calling`, `Undiplomatic Assertiveness / Whataboutism`, `Scapegoating`, `Propaganda Slinging`, `Appeal to Fear`, `Demonization`, `Personal Attacks`, `Doubt`, `Reductio Ad Hitlerum`, `Loaded Language`, `Appeal to False Authority` y `Bandwagoning`.

Como parte del leaderboard, el dataset tiene varias características interesantes:

- Se trata de un problema complejo (especialmente al nivel de caracterización de grano fino), difícil de resolver por cualquier sistema de PLN. Frente a otros datasets, éste tiene la ventaja de que no saturará fácilmente (momento en el cual no sirve para medir diferencias entre sistemas).
- Se trata de un problema de clasificación jerárquico, multiclase y multietiqueta. Aunque este tipo de problemas es muy habitual en aplicaciones prácticas de la IA, en entornos de laboratorio suelen simplificarse ignorando las características jerárquicas de las clases, o reduciendo el problema artificialmente a una sola etiqueta por ítem.
- Por tratarse de fuentes diplomáticas, abarca una gran variedad de registros dialectales, desde los propios del pais donde trabaja cada diplomático hasta su grado de bilingüismo (hay diplomáticos que son nativos en el idioma del pais de destino, otros lo han adquirido como segunda lengua).
- Se trata de una anotación de expertos, más costosa que el crowdsourcing pero que permite el uso de tipologías de anotación más sofisticadas, como es el caso de DIPROMATS.

Los detalles del dataset DIPROMATS 2023 están recogidos en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español: Informe del dataset DIPROMATS", entregado en el Año 1.

Dataset 3: DIANN 2023

El dataset DIANN-2023 se ha compilado y anotado en la UNED en el marco del proyecto. Consiste en un dataset bilingüe de resúmenes de artículos científicos relacionados con enfermedades raras anotados manualmente con discapacidades. El dataset ha sido concebido con el objetivo de entrenar sistemas de reconocimiento de entidades nombradas especializados en la detección de discapacidades. Una parte del corpus se creó en 2018 para la competición de IberLEF “Disability annotation on documents from the biomedical domain (DIANN)” [26](Fabregat et al., 2018), que proponía una tarea de reconocimiento de entidades centrada en la identificación de discapacidades. Otra parte se ha creado en 2023 con el fin de incorporarla como partición privada de test al leaderboard del proyecto ODESIA. La anotación de esta última parte es la que ha sido financiada por el proyecto.

El corpus se proporciona en dos particiones, una de entrenamiento y otra de evaluación. La partición de entrenamiento contiene 500 textos en cada lengua. Estos textos se corresponden con las particiones de entrenamiento y evaluación hechas públicas para la competición DIANN en Iberlef 2018, donde se proporcionaban 400 archivos de entrenamiento y 100 de evaluación por lengua. Además se dispone de una partición privada de test que contiene 100 textos para cada lengua. Puesto que esta es la partición que se usa para evaluar sistemas en el leaderboard, esta partición no se hará pública y no se proporciona información sobre sus contenidos más allá de la información referente al tamaño y la metodología de anotación.

En la Tabla 13 se muestran los detalles de tamaño para ambas particiones y el número anotaciones en la partición de entrenamiento. Los tokens se han calculado contando unidades separadas por espacios en blanco.

En el corpus se han anotado todas las discapacidades mencionadas en los textos. Por tanto, la única etiqueta usada es la de ‘discapacidad’. La distribución por lengua en la partición de entrenamiento es la siguiente:

[26]Página web de la competición DIANN 2018: http://nlp.uned.es/diann/

Partición	Entrenamiento				Evaluación		
Idioma	**# docs**	**# tokens**	**# líneas**	**# discapacidades**	**# docs**	**# tokens**	**# líneas**
Español	500	98948	5923	1555	100	21103	1203
Inglés	500	89325	6901	1656	100	19087	1234

Tabla 13: Información sobre el tamaño del corpus DIANN-2023.

- **Español**: 1555 menciones de discapacidades anotadas.
- **Inglés**: 1656 menciones de discapacidades anotadas.

Dentro del leaderboard, este dataset es representativo de uno de los dominios de aplicación más relevantes del PLN, el dominio biomédico; y de las tareas de etiquetado, que son las más prototípicas del PLN discriminativo junto con las de clasificación.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español, Informe del dataset DIANN-2023", entregado en el Año 1.

Dataset 4: EXIST-2023

En la línea de EXIST 2022, EXIST (sEXism Identification in Social neTworks) 2023 es un dataset desarrollado para facilitar la investigación en detección automática de sexismo en redes sociales. Se compone de textos cortos, tuits, procedentes de redes sociales etiquetados en función del tipo de sexismo expresado o descrito en ellos. A diferencia de EXIST 2022, se trata de un dataset desarrollado siguiendo el paradigma de "aprendizaje con desacuerdo" (Learning with Disagreements, LeWiDi) (Uma et al., 2021a), lo que lo convierte en el primer dataset para el entrenamiento y prueba de sistemas de detección de sexismo en textos construido conforme a este paradigma, y el primer dataset de este tipo que se incorpora al Leaderboard ODESIA.

En este paradigma, en lugar de depender de una única etiqueta "correcta" para cada ejemplo o instancia del dataset, el modelo se entrena para aprender de anotaciones conflictivas o diversas. De esta manera, las perspectivas, sesgos o interpretaciones diferentes de los anotadores pueden ser tenidas en cuenta por los sistemas, permitiendo un aprendizaje más justo y equitativo. Esto se deriva del hecho de que el dataset ha sido anotado por diferentes anotadores, con distintas características sociodemográficas, de manera que todas las notaciones (aun siendo en algunos casos contradictorias) forman parte del dataset final.

El dataset se compone de 10,034 textos etiquetados, 5,307 en español y 7,727 en inglés. Los textos proceden de conversaciones de Twitter. Cada texto está anotado con diferentes tipos de etiquetas. Como en EXIST 2022, en EXIST 2023 la primera etiqueta responde a la pregunta: *¿Es el texto sexista, en cualquiera de sus formas, o describe conductas o situaciones en las que se produce discriminación sexista (es decir, es sexista en sí mismo, describe una situación sexista o critica un comportamiento sexista)?* y la clasificación es binaria. A diferencia de EXIST 2022, en EXIST 2023 la segunda etiqueta responde a la pregunta: *¿Cuál crees que es la intención de la persona que escribió el tweet?* Las categorías anotadas son: `DIRECT`, `REPORTED`, `JUDGEMENTAL`. La tercera etiqueta responde a la pregunta: *De acuerdo con la faceta de la mujer que está siendo atacada, ¿qué tipo de sexismo encontramos en el tweet?* y las categoríaas son las mismas que en EXIST 2022.

El dataset presenta tres particiones para cada lengua: entrenamiento, desarrollo y test. La distribución de los textos por partición e idioma se muestra en la Tabla 14.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Año 2 - Informe del dataset EXIST 2023", entregado con este informe.

	Entrenamiento	Desarrollo	Test	Total
Español	3,660	549	1,098	5,307
Inglés	3,260	489	978	7,727
Total	6,920	1,038	2,076	10,034

Tabla 14: Distribución de EXIST 2023 por partición e idioma.

Dataset 5: SQUAD/SQAC-2024

En este caso, la tarea que este dataset permite evaluar es la de comprensión de texto en sistemas de pregunta-respuesta con respuestas extractivas. La tarea consiste en responder a preguntas sobre un texto, de tal manera que la respuesta sea un fragmento extraído directamente del texto. Los textos son noticias del CSIC para español y de Cambridge University para inglés. En todos los casos, las respuestas son fragmentos del texto y no se incluyen preguntas que no se puedan contestar a partir del texto. Esta tarea es interesante por su dificultad, ya que requiere tanto entender el lenguaje como tener una representación del conocimiento del mundo en general y del mundo representado en cada texto.

SQUAD/SQAC-2024 es una extensión de los datasets SQUAD/SQAC. SQAC (Spanish Question Answering Corpus) (Gutiérrez-Fandiño et al., 2021) es un dataset de pregunta-respuesta, con respuestas extractivas en español. En la tarea de PLN asociada, dada una pregunta y un párrafo, el sistema debe localizar el span (fragmento) más pequeño que contiene la respuesta. La metodología para crearlo está basada en la de SQuAD (Stanford Question Answering Dataset) v1.1 (Rajpurkar et al., 2016b), un dataset de pregunta-respuesta extractivo en inglés. Si bien el dataset SQAC se creó por la necesidad de tener un corpus de pregunta-respuesta en español que no fuera una traducción del inglés, el dataset SQUAD/SQAC-2024 se ha creado para disponer de una partición de evaluación privada que permita evaluar Large Language Models (LLMs) sin riesgo de contaminación de datos, tanto en inglés como en español.

El dataset contiene noticias académicas del CSIC (Centro Superior de Invvestigaciones Científicas) para el español[27] y de Cambridge University para el inglés.[28] Las noticias son de dominios científicos variados y suelen ser cortas, entre 712 y 2,760 palabras en inglés, y entre 514 y 2,818 palabras en Español. Además, están dirigidas al público general, por lo que no se usa lenguaje especializado.

En la Tabla 15 se proporciona información sobre el tamaño del dataset por idioma.

	# textos	# tokens	μ tokens/texto	# pares pregunta-respuesta	μ preguntas/texto
Español	110	962,502	840	1,144	10,4
Inglés	110	1,235,638	1,045	1,182	10,7
Total	220	2,198,140	–	2,379	–

Tabla 15: Información sobre el tamaño del dataset por idioma.

Las unidades que conforman el dataset son pares de pregunta-respuesta. Hay 1,144 pares en español y 1,182 pares en inglés, realizados sobre 110 textos en cada lengua. Los textos en español son un poco más cortos, con una media de 840 palabras, en comparación con los textos en inglés, con una media de 1,045 palabras. Se ha realizado una media de aproximadamente 10 preguntas por texto tanto en inglés como en español.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Año 2 - Informe del dataset SQUAD/SQAC-2024", entregado con este informe.

Datasets públicos adicionales para la medición de la brecha

[27] https://www.csic.es/es/actualidad-del-csic/noticias
[28] https://www.cam.ac.uk/news

Para afinar la brecha de efectividad inglés/español hemos utilizado todos los datasets bilingües disponibles en el Leaderboard ODESIA EXTENDED, que incorpora cuatro datasets bilingües adicionales construidos en inglés y español de dominio público. Hemos aplicado con ellos la misma metodología que con los de ODESIA CORE. Los datasets adicionales son:

- **DIANN Task 2** La segunda tarea de DIANN consiste en localizar y determinar el alcance de las negaciones en resúmenes de artículos biomédicos.
- **MLDoc** El Multilingual Document Classification Corpus (MLDoc) (Schwenk and Li, 2018) es un dataset de clasificación de noticias en varios idiomas, del que usamos el subconjunto de inglés y el subconjunto de español. La tarea tiene cuatro categorías: corporate/industrial, economics, government/social y markets.
- **MultiCONER 2022** (Malmasi et al., 2022) es un dataset multilingüe para reconocimiento de entidades nombradas complejas de seis categorías diferentes. Como en los demás casos, utilizamos los subconjuntos de español e inglés para nuestra experimentación. En este dataset, el conjunto de entrenamiento es un orden de magnitud más pequeño que el conjunto de evaluación (del orden de 5,000 vs 50,000).
- **STS-2017** (Cer et al., 2017b) es un dataset multilingüe de similitud textual, en la que los sistemas deben predecir el grado de similitud entre un par de oraciones. Esta es la única tarea de regresión en nuestro diseño experimental, y también la única en la que se usa como métrica de evaluación la correlación Pearson entre las predicciones del sistema y las anotaciones manuales. De nuevo, utilizamos los subconjuntos de inglés y español.
- **SQUAD/SQAC**: en este caso se trata de dos datasets construidos de forma independiente pero con la misma metodología. SQAC (Spanish Question Answering Corpus) (Gutiérrez-Fandiño et al., 2021) es un dataset de Question Answering extractivo, en el que, dada una pregunta y un párrafo asociado, el sistema debe localizar el span más pequeño que contiene la respuesta. La metodología para crearlo está basada en la de SQuAD v1.1 (Rajpurkar et al., 2016b), por lo que se pueden considerar datasets equivalentes. De hecho, nuestro algoritmo baseline obtiene un rendimiento equivalente en ambos idiomas (0,53 en ambos casos), lo que hace el par SQAC/SQUAD un conjunto ideal para comparar modelos de lenguaje en ambos idiomas. Nótese que este dataset también se utiliza como conjunto de entrenamiento en el SQUAD/SQAC 2024, pero en este último el conjunto de test se ha desarrollado en ODESIA sobre un tipo de documentos ligeramente diferente.

Para todos estos datasets utilizamos los datos públicos de entrenamiento y test, a diferencia de los datasets del Leaderboard ODESIA, en los que los conjuntos de evaluación no están disponibles públicamente. El proceso de entrenamiento para las tareas extendidas es idéntico al de las tareas de ODESIA CORE. En conjunto, la medición del gap se realiza sobre quince tareas.

5.3.3. Datasets adicionales no incorporados en el Leaderboard

Adicionalmente, se ha desarrollado los siguientes datasets que no se incorporan al Leaderboard público.

Dataset 6: CURIA-2024

CURIA-2024 es un dataset compuesto por sentencias de tribunales en inglés y en español, con sus correspondientes resúmenes simplificados. Se enmarca, por tanto, en el dominio jurídico. El dataset se ha elaborado con textos descargados de la página web del Tribunal de Justicia de la Unión Europea. Este organismo está formado por el Tribunal General (de primera instancia) y el Tribunal de Justicia. Todas las sentencias de ambos tribunales son publicadas en sus páginas web.[29] Aparte de las sentencias, CURIA-2024 contiene los micro-résumenes de las sentencias, que han sido creados por profesionales expertos en lenguaje legal.

[29] https://curia.europa.eu/jcms/jcms/j_6/es/

Puesto que el dataset se compone de sentencias con sus resúmenes, la unidad de referencia es el par texto/resumen. Como se muestra en la Tabla 16, el dataset en inglés contiene 2,230 pares texto/resumen sumando un total de 18,153,858 tokens, mientras que el dataset en español contiene 1,961 pares sumando un total de 17,842,388 tokens. El dataset está dividido en tres particiones por lengua: entrenamiento, desarrollo y evaluación. Las particiones se han creado siguiendo el mismo procedimiento para las dos lenguas: se ha tomado un 80 % para entrenamiento, un 10 % para desarrollo y un 10 % para evaluación.

Partición	**Entrenamiento**		**Desarrollo**		**Evaluación**		**Total**	
Idioma	**# docs**	**# tokens**	**# docs**	**# tokens**	**# docs**	**# tokens**	**# docs**	**# tokens**
Español	1,568	14,173,589	196	1,820,298	197	1,848,501	1,961	17,842,388
Inglés	1,784	14,461,442	223	1,724,975	223	1,967,441	2,230	18,153,858

Tabla 16: Número de pares sentencia–resumen en el dataset CURIA-2024.

Mediane este dataset se pueden evaluar sistemas de resumen simplificado de textos legales, por lo que se trata de una tarea de generación de texto.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español Informe del dataset CURIA-2024", entregado con este informe.

Dataset 7: UNED ACCESO 2024

Muchos benchmarks se han propuesto como evaluaciones de una sola tarea, pero con el surgimiento de modelos de lenguaje generales como BERT (Devlin et al., 2018), se ha popularizado el desarrollo de benchmarks más completos para poder medir las capacidades generales de estos modelos. GLUE (Wang et al., 2018a) y SuperGLUE (Wang et al., 2019) son benchmarks populares que evalúan el rendimiento de los modelos de lenguaje con distintas tareas de PLN. Más recientemente se han introducido benchmarks que contienen una amplia gama de tareas de PLN para la evaluación, como Big-Bench (Srivastava, 2022), Big-Bench Hard (Suzgun et al., 2022), MMLU (Hendrycks et al., 2021) y HELM (et al, 2023). La mayoría de los datasets presentes en estos benchmarks proponen evaluaciones que se realizan con conjuntos de datos creados artificialmente para tareas concretas de PLN, en vez de proponer escenarios reales de evaluación como los exámenes con los que se evalúa a los humanos. El reconocimiento de esta limitación ha llevado a que en los últimos tiempos se haya puesto énfasis en la importancia de las evaluaciones centradas en evaluar capacidades humanas. Se han introducido así benchmarks como AGIEval (Zhong et al., 2023), que se centran en un tipo de evaluación que pone el énfasis en tareas cognitivas a nivel humano, en escenarios reales. Este benchmark incluye exámenes de acceso a la universidad, pruebas de admisión a la facultad de derecho, concursos de matemáticas y pruebas de cualificación de abogados. Por otro lado, se han desarrollado diversos datasets a partir de exámenes, que abarcan distintas tareas (no sólo de respuesta múltiple), como en RACE (Lai et al., 2017). Las preguntas de respuesta múltiple se han erigido como uno de los métodos preferidos para evaluar los nuevos modelos generativos, debido a la dificultad que presenta su evaluación y la ausencia de una métrica estándar.

Así, en la intersección entre las evaluaciones con preguntas de múltiple respuesta, y las evaluaciones centradas en capacidades humanas y basadas en exámenes, se enmarca el dataset EXÁMENES UNED 2024. El dataset se compone de preguntas tipo test con 3 o 4 respuestas extraídas de exámenes de los Cursos de Acceso para Mayores de 25 años de la UNED de los siguientes grados: Administración y Dirección de Empresas, Biología, Bioquímica, Economía, Fundamentos de Informática, Lengua Castellana, Literatura, Matemáticas, Matemáticas Aplicadas a las Ciencias Sociales, Matemáticas Avanzadas y Psicología. El dataset se presenta en español y en inglés. La versión en español se ha obtenido directamente de la transcripción de los exámenes de Acceso para mayores de 25 años, mientras que la parte en inglés se ha construído mediante la traducción (manual) de estos exámenes.

En la Tabla 17 se incluye el número de preguntas, el número de exámenes y el número de palabras por asignatura, así como el número de opciones en la respuesta que tienen los exámenes de cada asignatura.

Los detalles sobre la metodología de construcción del dataset y sus características se encuentran en el informe técnico correspondiente, "Proyecto Espacio de Observación de Inteligencia Artificial en Español

Asignatura	# Preguntas	# Exámenes	# Respuestas por pregunta	# Palabras
Administración y Dirección de Empresas	87	6	3	3936
Biología	119	6	3	2872
Bioquímica	59	4	3	1466
Economía	51	3	4	1726
Fundamentos de Informática	63	6	4	1987
Lengua Castellana	94	4	4	2816
Literatura	91	6	4	5130
Matemáticas	73	11	3	1538
Matemáticas Aplicadas a las Ciencias Sociales	94	10	3	2941
Matemáticas Avanzadas	24	5	3	470
Psicología	248	14	4	5669
Total	1003	75	–	30551

Tabla 17: Distribución del número de preguntas y exámenes por asignatura, y número de opciones de respuesta por pregunta que tienen los exámenes de cada asignatura.

Informe del dataset UNED ACCESO 2024", entregado con este informe.

Dataset 8: PRON VS PROMPT

Motivación Históricamente, hitos en el desarrollo de la Inteligencia Artificial como las victorias de Deep Blue y AlphaGo en ajedrez y Go, respectivamente, han marcado el avance tecnológico en diversas área de investigación. En el caso del Go, el sistema de IA desarrolló estrategias de juego novedosas que han sido imitadas, desde entonces, por todos los maestros humanos: se demostró que la IA podía ser creativa. Pero los juegos de mesa tiene características muy particulares que los hacen muy adecuados para los sistemas de IA. Desde la aparición de ChatGPT, la atención se ha desplazado hacia tareas como la escritura creativa, mucho más complejas. Estos modelos de lenguaje desafían la concepción de la inteligencia humana y las fronteras de la creación artística tradicionalmente considerada exclusiva de los humanos. Tanto es así que la colaboración humano-máquina en la industrias creativas es cada vez mayor (Adelani et al., 2023). De hecho, tal es su relevancia que los propios desarrolladores de OpenIA, en la interfaz de ChatGPT, ofrecen como primera sugerencia de uso de su modelo de lenguaje la opción çrea un historia".

Objetivo A pesar de todo, no existen aproximaciones de evaluación objetiva y rigurosa en esta tarea. Este dataset tiene el propósito general medir comparativamente la calidad en la creación de textos creativos de GPT4, el modelo con mejor rendimiento hasta la fecha, y un escritor profesional. Así, el objetivo de este dataset es doble: por un lado, se quiere realizar un estudio comparativo entre las capacidades creativas de generación de texto de una Inteligencia Artificial avanzada (GPT-4) y un novelista consagrado, Patricio Pron (Premio Alfaguara de novela). Y, por otro lado, se diseñará una metodología rigurosa de evaluación de textos creativos que servirá para puntuar la salida de cualquier modelo. A medio plazo, esperamos que esta experimentación sea el punto de partida para el desarrollo de métricas de evaluación automática que permitan conocer de la manera más objetiva posible la calidad de las producciones literarias de los modelos de lenguaje.

Descripción El dataset consiste en 120 sinopsis para 60 títulos de películas imaginarias, 30 propuestos por GPT4 y 30 por Patricio Pron. Se solicitó a ambos, al escritor y a GPT-4, que escribieran sinopsis de aproximadamente 600 palabras para cada título, incluyendo tanto los propuestos por ellos mismos como por su contrincante.

Evaluación Los textos generados están (a fecha de finalización del año 2 del proyecto) siendo sometidos a una evaluación a ciegas por un panel de expertos, compuesto por críticos y académicos, para garantizar

una valoración objetiva de la calidad, creatividad y coherencia narrativa de las sinopsis. Los resultados estarán disponibles a lo largo del tercer año del proyecto.

5.3.4. Modelos de lenguaje utilizados en la experimentación

Para establecer una comparación inicial entre modelos del lenguaje en inglés y español, hemos hecho una selección de modelos preentrenados en español, en inglés, y multilingües, escogiendo entre los más utilizados en la literatura y en Hugging Face. En conjunto se trata de cinco modelos en español, cuatro en inglés, y cinco modelos multilingües, lo que nos permite considerar diez modelos en el leaderboard español y nueve en inglés.

El leaderboard incluye tareas tanto discriminativas (EXIST, DIPROMATS, etc.) como generativas (CURIA, UNED Acceso). De manera natural, las tareas discriminativas suelen resolverse con modelos encoders. Y, las tareas generativas, con modelos decoder, también llamados modelos generativos; el objetivo de entrenamiento de los modelos decoder es generar texto fluido. Sin embargo, para resolver esta tarea, han demostrado adquirir una gran capacidad de comprensión de lenguaje tal que parecen ser capaces de resolver tareas discriminativas, aunque no estén originalmente diseñados para eso.

De momento en el leaderboard sólo se incluyen los modelos discriminativos. Respecto a modelos generativos, en este año se ha realizado una primera experimentación con el dataset UNED-ACCESO, utilizando los modelos decoder o generativos que ahora mismo son el estado del arte: GPT4 (Achiam et al., 2023), GPT3.5[30], Claude 3[31], Mistral (Jiang et al., 2023), Llama-2 (Touvron et al., 2023) y Gemma[32]. Además, está en proceso la evaluación de estos modelos con el dataset CURIA de resúmenes en texto claro. También se ha evaluado GPT-4 en modo zero-shot (proporcionando al sistema sólo la guía de anotación) para las tres tareas de DIPROMATS 2023.

Los modelos discriminativos incluidos en el leaderboard son los siguientes:

Español	Inglés	Multilingües
roberta-large-bne	roberta-large	xlm-roberta-large
roberta-base-bne	roberta-base	xlm-roberta-base
bert-base-spanish-wwm-cased	bert-base-cased	bert-base-multilingual-cased
distillbert-base-spanish-uncased	distilbert-base-uncased	distilbert-base-multilingual-cased
bertin-roberta-base-spanish		ixambert-base-cased

Tabla 18: Modelos de lenguaje evaluados.

Modelos en inglés

- **Bert-base-cased** BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) fue el primer modelo que utilizó la arquitectura transformer, que después ha sido la predominante en toda la investigación posterior. El modelo base tiene 12 capas transformer blocks, 12 attention heads, y 110 millones de parámetros. El preentrenamiento se realizó de forma autosupervisada en un gran corpus en inglés optimizando dos objectivos simultáneamente:
  - Masked Language Model (MLM). En cada frase el modelo oculta de forma aleatoria el 15 % de las palabras, e intenta predecir las palabras ocultas.
  - Next Sentence Prediction (NSP). En el preentrenamiento, dado un par de oraciones, el modelo debe predecir si las dos frases aparecian consecutivamente en el corpus de entrenamiento o no.
- **RoBERTa-base** RoBERTa (Liu et al., 2019) un modelo que introduce algunas variaciones respecto a BERT. La principal diferencia es que RoBERTa usa un masking dinámico, en el que en cada epoch (iteración de aprendizaje sobre el corpus de entrenamiento) se enmascaran distintas partes de la oración. RoBERTa-base tiene 123M de parámetros.

[30] https://chat.openai.com
[31] https://www.anthropic.com/news/claude-3-family
[32] https://ai.google.dev/gemma?hl=es-419

- **Roberta-large** La única diferencia entre ROBERTA-BASE y ROBERTA-LARGE es el número de parametros utilizado para su entrenamiento. Si bien ROBERTA-BASE fue entrenado con 123 millones de parámetros, ROBERTA-LARGE lo hizo con 354 millones de parámetros.
- **Distilbert-base-uncased** es un modelo preentrenado con el mismo corpus que BERT, pero se trata de un modelo más pequeño y más rápido (Sanh et al., 2019). El modelo fue preentrenado con tres objectivos:
  - Que tuviera la capacidad de devolver las misma probabilidades que el modelo BERT
  - Que fuera capaz de aprender como su predecesor al ocultar el 15 % de las palabras en la entrada.
  - Que fuera capaz de generar estados ocultos como su predecesor.

  Con esto se pretendía conseguir un modelo igual de potente que BERT, pero a su vez más rápido y pequeño.

Modelos en español

- **Roberta-base-bne** Este es un modelo basado en RoBERTa base que ha sido preentrenado con el mayor conjunto de datos disponible en español hasta la fecha de su entrenamiento. Forma parte del conjunto de modelos MarIA (Gutiérrez-Fandiño et al., 2021) desarrollados por el Barcelona Supercomputing Center y la Biblioteca Nacional de España dentro del Plan Nacional de Tecnologías del Lenguaje español. El corpus de entrenamiento está compuesto por un total de 570 GB de texto limpio, realizado por la Biblioteca Nacional de España a partir del rastreo de páginas web entre 2009 y 2019. Sin embargo, la cantidad de datos para entrenar roberta-base-bne fue inferior, para obtener un modelo más ligero.
- **Roberta-large-bne** Modelo RoBERTa large entrenado de forma similar al anterior.
- **Bertin-roberta-base-spanish** Este es un modelo entrenado desde cero sobre la porción en español de mC4 (Xue et al., 2020), que contiene unos 416M de documentos. Se trata de un modelo RoBERTa-base con 12 capas, 12 attention heads cada una, y 125M de parámetros.
- **Bert-base-spanish-wwm-cased**, conocido como *BETO* (Cañete et al., 2020), es un modelo entrenado en un gran corpus español y que tiene un tamaño similar a BERT-Base. Al igual que en las versiones anglosajonas, esta versión distingue entre mayúsculas y minúsculas.
- **Distillbert-base-spanish-uncased** Se trata de la versión "destilada"de BETO. Fue entrenada por sus mismos autores y tiene el propósito de ofrecer un modelo de un tamaño más reducido que BETO, pero con un rendimiento similar.

Modelos bilingües

- **xlm-roberta-base** (Conneau et al., 2019) Se trata de la versión multilingüe de RoBERTa, abarca más de 100 idiomas y, a pesar de que en la página de Huggingface asegura que está preentrenada con 2,5TB de datos CommonCrawl, está entrenado con menos cantidad de datos que la versión *large*.
- **xlm-roberta-large** Esta es la versión que sí utiliza toda la cantidad de datos de los 2,5TB. Al igual que el modelo inferior, su entrenamiento está realizado para más de 100 idiomas diferentes.
- **bert-base-multilingual-cased** Modelo BERT preentrenado con los 104 idiomas mejor representados en la Wikipedia.
- **distilbert-base-multilingual-cased** Modelo DistilBERT preentrenado con un corpus similar al modelo anterior.
- **ixa-ehu/ixambert-base-cased** (**?**) Modelo preentrenado multilingüe para inglés, español y euskera. El corpus de entrenamiento está compuesto por las Wikipedias en inglés, español y euskera, junto con artículos de noticias en euskera extraídos de periódicos online.

Entrenamiento de modelos e hiperparámetros

El entrenamiento eficaz de modelos de lenguaje avanzados es fundamental para que den su mejor rendimiento. Un aspecto fundamental de este proceso es la selección óptima de hiperparámetros, los cuales pueden influir significativamente en la capacidad del modelo para aprender de los datos. En este contexto, hemos implementado una estrategia de búsqueda exhaustiva, conocida como Grid Search, para optimizar los hiperparámetros de nuestros modelos de lenguaje.

Grid Search es una técnica de optimización de hiperparámetros que sistematiza el proceso de experimentación al construir y evaluar un modelo para cada combinación de parámetros especificada en una cuadrícula predefinida. Esto permite identificar la configuración de hiperparámetros que resulta en el mejor rendimiento del modelo.

La aplicación de Grid Search asegura una exploración exhaustiva del espacio de hiperparámetros, proporcionando una base sólida para la toma de decisiones informadas sobre la configuración óptima para el entrenamiento de modelos.

Para la optimización de los modelos de lenguaje, hemos definido la siguiente cuadrícula (grid) de hiperparámetros para explorar:

- Tamaño de *batch*: Especifica el tamaño del lote (batch size) durante el entrenamiento. Se consideraron tamaños de 32 y 16, ajustando el equilibrio entre la memoria consumida y la precisión del gradiente.
- Tasa de aprendizaje: Se utiliza para el ajuste de los pesos del modelo en cada iteración. Se exploraron valores de 0.00001, 0.00003 y 0.00005, buscando optimizar la velocidad y estabilidad del aprendizaje.
- Weight decay: Controla la regularización sobre los pesos del modelo. Se usaron los valores de 0.1 y 0.01. La regularización ayuda a prevenir el sobreajuste al penalizar pesos grandes.

Para cada combinación de hiperparámetros especificada en la grid, se entrenó un modelo de lenguaje desde cero, utilizando un conjunto de datos estandarizado. La evaluación del rendimiento de cada modelo se realizó a través de las métricas específicas que se detallan en la siguiente Sección, tratando de identificar la configuración de hiperparámetros que maximiza el rendimiento del modelo sobre el conjunto de desarrollo. Cada modelo se entrenó un total de 12 veces en diferentes configuraciones sobre cada tarea e idioma.

La aplicación del Grid Search ha permitido una exploración sistemática y completa del espacio de hiperparámetros, identificando la configuración que conduce al mejor rendimiento de los modelos de lenguaje en el contexto de nuestro proyecto. La metodología adoptada asegura que la selección de hiperparámetros se base en evidencia empírica, mejorando la confiabilidad y eficacia de los modelos entrenados, así como la fiabilidad del gap reportado.

Hemos realizado una configuración base para todos los modelos, con alguna alteración para los problemas de clasificación binaria y multiclase, ya que los modelos DistillBert no soportan el *'gradient_checkpoint'*. Sin embargo, algunos modelos requieren tener activada está opción debido a la carga de memoria que reciben a la hora de procesar la información en la capa de *Attention*, sobre todo en modelos grandes con muchos parámetros y en modelos que no convergían por el número de pasos o epochs. Según el conjunto de datos y sus tareas, hemos ajustado diferentes configuraciones para comprobar la consistencia de los modelos. En cualquier caso, la configuración final para cada tarea se estableció como única para todos los modelos.

Evaluación: Métricas y baselines

En esta sección resumimos las métricas y los algoritmos baseline, sin información lingüística, utilizados como referencia para calibrar la efectividad de los modelos del lenguaje entre inglés y español.

Para poder comparar resultados en datasets de dos idiomas es necesario estimar primero la dificultad intrínseca de cada dataset, de forma que puedan calibrarse los resultados en un idioma y otro para compararlos directamente. Para ello, en cada dataset hemos estimado la dificultad intrínseca mediante algoritmos de Machine Learning que no usan ningún tipo de información lingüística.

Respecto a las métricas, salvo que se indique lo contrario se ha utilizado la implementación de las métricas en la librería PyEvall desarrollada dentro del proyecto.

EXIST-2022

Métrica Para evaluar las dos tareas de EXIST-2022, se ha utilizado F1 Macro (ver informe correspondiente).

Baseline Para generar los baselines de las tareas de EXIST, vectorizamos el conjunto de datos y test sin ningún tipo de preprocesamiento, para evitar utilizar información lingüística. A partir de los conjuntos resultantes, entrenamos modelos de regresión logística, xgboots y Support Vector Machines (SVM). Tomamos como baseline la media de los resultados obtenidos. El proceso es el mismo para las dos tareas.

DIANN 2023

Métrica La métrica que hemos usado para evaluar la tarea de DIANN 2023 es F1, para ello hemos usado Seqeval,[33] un framework en python que evalúa el etiquetado de secuencias.

Baseline Para realizar la tarea de NER de DIANN, utilizamos Conditional Random Fields (CRF), una clase de métodos de modelado estadístico que se aplican a menudo en el reconocimiento de patrones. Lo hemos usado como fórmula de predicción estructurada. No se ha usado ningún tipo de información lingüística.

DIPROMATS 2023

Métrica La evaluación se ha llevado a cabo considerando las tareas como de clasificación, abarcando tanto clasificación binaria, para la Tarea 1, como multietiqueta, para las Tareas 2 y 3. Es importante destacar que la tarea de clasificación multietiqueta de estas dos últimas presenta una complejidad no trivial desde el punto de vista de las métricas de evaluación. Esto se debe a que las clases involucradas mantienen una relación jerárquica entre sí. Por ejemplo, en la Tarea 2, un error de clasificación entre el grupo 2 y el grupo 3 se considera menos grave que un error entre el grupo 2 y el grupo 0.

Por ello, además de las métricas estándar, se reporta la métrica ICM (Amigó and Delgado, 2022), la cual se adapta de manera óptima a las particularidades de nuestra tarea de clasificación. La métrica ICM está diseñada para considerar con severidad variable los errores de clasificación entre diferentes grupos, basándose en su relación jerárquica. Esto la hace especialmente adecuada para tareas en las que los errores entre ciertas clases son inherentemente menos críticos que entre otras, como en DIPROMATS.

Aparte, ya que son métricas estándar, se reportan la Precisión (P), Recall (R) y la puntuación F1 (F1) de cada clase, para proporcionar una evaluación básica del rendimiento de los modelos en las tareas de clasificación mencionadas.

Baseline Para la primera tarea de DIPROMATS, usamos los mismos algoritmos que hemos usado en la clasificación binaria de EXIST 2022. Para las tareas 2 y 3 de DIPROMATS, y continuando con la idea de evitar usar información semántica en los modelos, usamos un clasificador multietiqueta, basado en el algoritmo de KNN (K-Nearest Neighbors).

EXIST-2023

Métrica La métrica empleada fue la oficial de la competición en el modo de evaluación soft-soft (Plaza et al., 2023), ICM-soft. ICM-soft se ha desarrollado dentro del proyecto ODESIA, y es una extensión de la métrica ICM (Amigó and Delgado, 2022) que permite evaluar un sistema bajo el paradigma "learning-with-disagreements"(LeWiDi) (Uma et al., 2021b) comparando sus salidas (dadas como probabilidades de pertenencia a una o varias clases) con un “soft gold standard” especificado de esta misma forma. Además, ICM-soft permite evaluar distintos tipos de problemas de clasificación: binaria (Tarea 1), jerárquica multiclase (Tarea 2) y jerárquica multiclase multietiqueta (Tarea 3). Es la única métrica existente que permite evaluar tareas complejas de clasificación (multilabel, jerárquica o ambas) en modo learning with

[33] https://huggingface.co/spaces/evaluate-metric/seqeval

disagreement.

Baseline para cada tarea Para establecer un baseline para cada una de las tareas de EXIST-2023 se entrenó una red neuronal simple con una única capa oculta para la clasificación de los textos. La red se entrenó a 20 epochs con un learning rate de 0.001. El algoritmo de optimización escogido fue Adam. La función de pérdida usada por la red neuronal fue la entropía binaria para las tres tareas. La arquitectura de la red neuronal constó de tres componentes principales: dos capas totalmente conectadas y una función de activación unitaria lineal rectificada (ReLu). Los textos de entrada suministrados a la red se conviertieron primeramente a vectores de 10.000 dimensiones mediante el método TF-IDF La última capa de la red neuronal produce salidas que corresponden al número de capas objetivo de cada tarea específica: 2 para la tarea 1, 4 para la tarea 2 y 6 para la tarea 3. Para obtener un baseline robusto en cada tarea, se promediaron los resultados de 10 ejecuciones distintas para cada tarea.

SQUAD/SQAC-2024

Métrica Para evaluar la tarea de SQAC-SQUAD-2024 hemos usado la misma métrica que se utilizó para evaluar tanto SQAC como SQUAD v1.1. Para calcular el F1 score, primero se hace un preprocesamiento de las predicciones y gold standard, luego cada par de respuestas (predicción-gold standard) se tokenizan y se cuenta cuantas palabras coinciden. Con este dato, se calcula el F1.

Baseline El algoritmo baseline consiste en cotejar, mediante distancia coseno cada frase del texto con la pregunta, tomando la más semejante como respuesta candidata.

MLDoc

Métrica Para la tarea de MLDoc, al ser una tarea de clasificación multi-categoría, hemos usado la métrica estándar F1 macro.

Baseline Para este conjunto de datos usamos los algoritmos ya comentados de Regresión logística, Xgboots y SVM. Al igual que en los anteriores baseline, vectorizamos evitando utilizar información lingüística.

MultiCONER 2022

Métrica Al igual que DIANN, MultiCONER es una tarea NER, por lo que hemos usado la misma métrica que para DIANN (F1 en la implementación del script Seqeval).

Baseline Para MultiCONER 2022, usamos el mismo algoritmo baseline que para DIANN, CRF.

STS 2017

Métrica STS es una tarea de similitud, y para medir esa similitud entre dos frase, usamos la métrica de Pearson Correlation.

La correlación de Pearson se calcula con la siguiente fórmula:

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

donde:
- r es el coeficiente de correlación de Pearson,
- x_i e y_i son los valores individuales de las variables X e Y,
- $\overline{x}$ y $\overline{y}$ son los promedios de las variables X e Y,
- $\sum$ denota la suma sobre todos los datos de muestra.

Baseline En el caso de STS 2017, un problema de regresión, simplemente vectorizamos las dos frases de cada caso mediante la función `TfidfVectorizer` de sklearn (Pedregosa et al., 2011), y calculamos el coseno entre ambas representaciones como aproximación de su similitud.

SQAC/SQUAD

Métrica Para evaluar la tarea de SQAC-SQUAD hemos usado una versión flexible de F1 (lenient frente a strict). Para calcular este F1 score, primero se hace un preprocesamiento de las predicciones y gold standard, luego, con cada par de predicción-gold standard, se tokenizan y se cuenta cuantas palabras coinciden. Con este dato de todos los pares, se calcula el F1.

Baseline El algoritmo baseline consiste en cotejar, mediante distancia coseno como en el caso anterior, cada frase del contexto con la pregunta, quedándose con toda la frase como respuesta candidata. Nótese que al tomar toda la frase como respuesta (para evitar cualquier tipo de proceso lingüístico), la puntuación del baseline es muy baja.

DIANN

Métrica La métrica que hemos usado para evaluar la tarea es F1, para ello hemos usado Seqeval, un framework en python que evalúa el etiquetado de secuencias

Baseline Para obtener el baseline de DIANN Detección de negación utilizamos Conditional Random Fields (CRF), una clase de métodos de modelado estadístico que se aplican a menudo en el reconocimiento de patrones. Lo hemos usado como fórmula de predicción estructurada. La idea, al igual que en las demás experimentaciones, es que no haya información semánticasen los baselines.

5.3.5. Resultados experimentales: gap en efectividad

En la Tabla 19 pueden verse los resultados de nuestra experimentación sobre las tareas del Leaderboard ODESIA EXTENDED. Disponemos de 15 mediciones del gap sobre cada una de las 15 tareas contempladas. En cada una de ellas, los baseline se han calculado como el promedio de varios algoritmos de aprendizaje supervisado sin conocimiento lingüístico, como se ha explicado anteriormente. Las columnas *mejor ES* y *mejor EN* recogen la efectividad del modelo de lenguaje con mejores resultados en cada tarea en español e inglés, respectivamente.

Es interesante observar que, a pesar de la diversidad de datasets y tareas, en todos los casos menos uno (EXIST 2023 tarea 2, donde el rendimiento en ambos idiomas es similar) se ha medido una brecha favorable al inglés, lo que proporciona una fuerte evidencia estadística de la existencia de esa brecha. Puede apreciarse, también, que hay un valor muy diferente al resto en el caso de la tarea DIPROMATS Task 3. Esa tarea es la más difícil de las contempladas en nuestra experimentación, ya que es un problema de clasificación multiclase y multilabel con 13 clases distribuidas de forma muy desigual. Al ser la más difícil, también es la tarea en la que los modelos de lenguaje aportan una mejora más sustancial respecto a las aproximaciones baseline sin conocimiento lingüístico. Aunque es un resultado en el que merece la pena profundizar, a efectos del cálculo del gap debemos descartarlo por razones estadísticas, ya que se trata de un outlier ($p < 0,01$ según el test de Grubbs[34]).

Una vez eliminado el outlier, el indicador de brecha de efectividad EF,1 se ha calculado según se describe en la sección 4.1.6 sobre las otras catorce tareas. **El resultado final (con su error estándar) es una estimación de la brecha porcentual del 20 ± 06 a favor del inglés.** Aunque hay una variación apreciable entre tareas, el error estándar nos indica que, en cualquier caso, la brecha real promedio estará en una horquilla entre el 14 % y el 26 %. Estos datos son compatibles con los obtenidos el primer año.

Esta brecha está medida exclusivamente sobre modelos discriminativos. Nuestra experimentación con modelos generativos a dado lugar a primeras estimaciones de brecha que no hemos incluido en el indicador global, por las razones que se comentan en la sección 5.4.

5.4. Evolución de la brecha de efectividad

En el primer año de proyecto se midió una brecha de efectividad de modelos discriminativos del 18 ± 04 %, y en este segundo año la medición es del 20 ± 06 %. Las dos mediciones son compatibles dentro de su error estándar y, por tanto, **no se puede hablar de variación** en este aspecto. De hecho, no han surgido nuevos modelos discriminativos destacables durante este año, de modo que la pequeña diferencia observada proviene de la mejora del diseño experimental, al que hemos añadido cuatro tareas nuevas y una búsqueda exhaustiva de hiperparámetros (con más de 3600 experimentos).

[34] Hemos utilizado https://www.graphpad.com/quickcalcs/grubbs2/

Tarea	baseline ES	Baseline EN	mejor ES	mejor EN	brecha %
Core Tasks					
EXIST-2022 Task 1	0,693	0,674	0,770	0,810	16,54
EXIST-2022 Task 2	0,463	0,437	0,570	0,580	9,803
DIPROMATS-2023 Task 1	0,750	0,707	0,818	0,819	11,60
DIPROMATS-2023 Task 2	0,220	0,210	0,471	0,550	47,80
DIPROMATS-2023 Task 3	0,090	0,080	0,267	0,472	293,33
DIANN Task 1	0,747	0,665	0,840	0,790	0,38
EXIST-2023 Task 1	0,469	0,437	0,645	0,643	9,52
EXIST-2023 Task 2	0,251	0,220	0,421	0,363	-2,70
EXIST-2023 Task 3	0,218	0,206	0,400	0,403	12,10
SQUAD/SQAC-2024	0.132	0.125	0,464	0,463	18,60
Extended Tasks					
DIANN Task 2	0,878	0,575	0,960	0,917	71,80
MLDoc	0,930	0,883	0,970	0,980	39,86
MultiCoNER 2022	0,523	0,553	0,710	0,750	4,86
STS-2017	0,680	0,707	0,800	0,860	14,76
SQUAD/SQAC	0,533	0,528	0,770	0,883	24,57
Brecha efectividad					**20 $\pm$ 06**

Tabla 19: Cálculo de la brecha de efectividad. Los baseline se han calculado como el promedio de varias algoritmos de aprendizaje supervisado sin conocimiento lingüístico. Los LLM son la efectividad del modelo de lenguaje con mejores resultados en cada tarea. El indicador de brecha de efectividad E,1.a se ha calculado según se describe en la sección 4.1.6, eliminando del promedio el outlier *DIPROMATS Task 3* y reportando el error cuadrático medio sobre el resto de mediciones. Un número positivo indica una brecha porcentual a favor del inglés, y negativo a favor del español.

La razón de que no hayan aparecido modelos discriminativos destacables es que el esfuerzo investigador (tanto en la academia como en la industria) se ha volcado en los modelos generativos. Estos modelos no son apropiados para tareas discriminativas para las que se tienen datos de entrenamiento (que son muchas de las que tienen aplicación industrial), pero han abierto un gran abanico de aplicaciones en modo no supervisado que eran impensables hace pocos años. Nuestra experimentación inicial con modelos generativos no tiene suficiente representatividad como para incluirla en el cálculo de la brecha, pero indica, provisionalmente, que **en los modelos generativos no hay una brecha mayor que la que hemos medido con los discriminativos**:

- GPT-4 (el modelo más potente junto con Claude 3 en el momento de finalizar nuestra experimentación) en modo zero-shot, aplicado a tres tareas discriminativas de nuestro leaderboard, arroja una brecha promedio del 18 % entre el español y el inglés, muy similar a nuestro resultado en modelos discriminativos.

- En nuestro dataset UNED ACCESO de preguntas de exámenes de acceso a la universidad, los modelos generativos abiertos LLama-2, Gemma y Mistral tienen una brecha promedio del 12 % entre el español y el inglés. Y los modelos más potentes (Claude-3 y GPT-4) dan una brecha ligeramente negativa (-2 % en ambos casos), es decir, se comportan un poco mejor en español que en inglés. Teniendo en cuenta que las preguntas del dataset están originalmente en español (y son traducidas manualmente al inglés), es muy posible que los modelos hayan visto parte de las respuestas en su fase de entrenamiento (contaminación), y tampoco es descartable que haya algún artefacto de traducción (aunque son traducciones manuales realizadas por profesionales). Descartando estos efectos, no

sería raro que la medición del gap estuviera en niveles parecidos a los del experimento anterior con GPT-4. En cualquier caso, necesitamos profundizar en nuestro diseño experimental para obtener cifras fiables.

6. Cálculo de indicadores: Ámbito 2 - Soluciones de mercado

En esta sección se presentan los resultados del cálculo de indicadores sobre soluciones de mercado de tecnologías de la lengua. En primer lugar, se presenta un análisis detallado de la relevancia de las funcionalidades de las soluciones de mercado seleccionadas. A continuación se realiza un análisis comparativo detallado de las funcionalidades y, por último, se presenta el cálculo del indicador de brecha en funcionalidades.

6.1. Análisis de relevancia de funcionalidades

En esta sección se analiza la relevancia de las funcionalidades para cada una de las áreas de aplicación.

6.1.1. Análisis de opiniones

La clasificación de sentimiento y clasificación de impacto reputacional son dos funcionalidades importantes para cualquier software de análisis de opinión o social listening. Con la clasificación de sentimiento, las empresas pueden identificar la opinión general de los usuarios sobre su marca y detectar mensajes positivos, negativos o neutros. Por otro lado, la clasificación de impacto reputacional permite a las empresas identificar los mensajes que pueden tener un impacto negativo en la reputación de la marca y tomar medidas preventivas para proteger su imagen.

La detección de emociones, temas de conversación y entidades también es relevante para comprender mejor las necesidades y deseos de los clientes. La detección de emociones puede ayudar a las empresas a conocer mejor el estado de ánimo de los clientes, mientras que la detección de temas de conversación puede ayudarles a detectar los temas que son importantes para sus clientes y adaptar su estrategia en consecuencia. La detección de entidades, como nombres de productos o marcas, permite a las empresas identificar las menciones de su marca y marcas de la competencia en las redes sociales y realizar una comparativa y seguimiento de su reputación online.

La detección de mensajes inapropiados es de interés para el filtrado de opiniones, mientras que la detección de motivaciones o estímulos de los emisores de los mensajes puede ser útil para comprender mejor las necesidades y deseos de los clientes. Finalmente, la posibilidad de definir las clases y ajustar el modelo son funcionalidades que pueden tener interés para adaptar la herramienta a las necesidades específicas de la empresa y obtener resultados más precisos y personalizados.

6.1.2. Asistentes virtuales

En los asistentes virtuales tradicionales, el reconocimiento de voz es una funcionalidad fundamental, ya que permite al asistente activarse, reconocer quién está hablando y responder de manera personalizada. La posibilidad de añadir habilidades también es importante, ya que permite al asistente interactuar con otros servicios y mejorar su funcionalidad. Los acentos o variantes regionales también son importantes en el caso de usar los asistentes con idiomas globales para asegurar una comunicación efectiva en diferentes regiones del mundo.

Los comandos aceptados son otra funcionalidad importante en un asistente virtual, ya que permiten al usuario realizar acciones específicas, cómo configurar una alarma o saber el tiempo que va a hacer. La capacidad de escribir texto es otra funcionalidad útil que permite al usuario dictar texto y traducirlo a texto escrito. Por último, otras funcionalidades, como el reconocimiento de entidades, stemming y autocorrección, también pueden ser importantes para mejorar la precisión y eficiencia del asistente virtual.

Por último, chatbots como ChatGPT están basados en modelos de lenguaje a gran escala y su principal funcionalidad es responder preguntas y generar texto coherente. A diferencia de los asistentes virtuales tradicionales, ChatGPT no está diseñado para realizar tareas específicas, sino para mantener conversaciones y proporcionar información relevante. Cabe señalar que el mapa de funcionalidades que se ha propuesto para este proyecto se definió antes de la irrupción de ChatGPT en la escena de la IA en noviembre, por lo

que se prevé adaptarlo para el segundo año, teniendo en cuenta funcionalidades específicas de este tipo de tecnología.

6.1.3. Traducción automática

En una herramienta de traducción automática la capacidad de traducir desde una gran cantidad de idiomas es vital si se quiere dar respuesta a las diferentes necesidades de los usuarios. La corrección gramátical y la detección del idioma de partida también son funcionalidades de interés, ya que permiten al usuario asegurarse de que la traducción es precisa. La capacidad de ofrecer diferentes traducciones para variantes regionales o personalización a dominios específicos también puede ser útil para obtener textos que se adapten a la realidad profesional y regional de los usuarios.

La posibilidad de modificar las traducciones, ofrecer diferentes versiones de la traducción y traducir archivos y páginas web completas son funcionalidades útiles que mejoran la experiencia del usuario. Por último, con la proliferación de los smartphones con cámaras y micrófonos, las fuentes de datos cada vez son de tipo más diverso y es importante que las soluciones no se centren únicamente en la traducción de texto a texto, sino también a la traducción de voz, imágenes o vídeos.

6.1.4. Teclados predictivos

En un teclado predictivo, la corrección gramátical es una funcionalidad importante que ayuda a mejorar la calidad de la escritura y a evitar errores que son muy comunes cuando se escribe con los teclados virtuales de los smartphones. La capacidad de personalizar la predicción del teclado a la manera de escribir del usuario y la capacidad de poder dictar los mensajes de voz a texto son importantes, ya que hacen que la experiencia de escritura sea más precisa y eficiente. La detección de idioma es otra funcionalidad útil que permite al teclado ajustar la predicción a la lengua de partida del usuario.

El autocompletado de palabras y las sugerencias de palabras también son importantes para acelerar la escritura y evitar errores. En este sentido, si el teclado predictivo puede sugerir expresiones, frases completas o textos más extensos como párrafos, puede resultar de gran utilidad para aquellos que necesiten escribir textos largos o artículos completos.

6.1.5. Buscadores web

En un buscador web, la corrección gramatical es una funcionalidad importante que ayuda a mejorar la calidad de la búsqueda. La capacidad de clasificar los documentos por tema es otra funcionalidad útil que permite al usuario filtrar los resultados de búsqueda por diferentes categorías, como el tipo de página o el tema del texto. La detección de entidades también es importante, ya que permite al buscador identificar nombres, sitios y empresas relevantes en los documentos. Asimismo, la búsqueda de sinónimos es otra funcionalidad que mejora la precisión de la búsqueda.

La red cada vez incluye más contenido audiovisual. En este sentido, la capacidad de buscar imágenes, texto en imágenes, vídeos y audio también es muy relevante para aquellos que necesitan buscar contenido multimedia. Por último, existen funcionalidades más profundas como la detección de significado y la capacidad de devolver respuestas. La primera permite buscar sobre el significado de la frase en lugar de limitarse a la coincidencia de palabras. La segunda devuelve la información de manera estructurada en lugar de limitarse a mostrar URLs de manera ordenada y, de este modo, mejora de forma notable la experiencia de usuario.

6.2. Comparativa de funcionalidades

En esta sección se realiza un análisis de las funcionalidades que ofrecen cada una de las soluciones. Cabe destacar, que el análisis se centra en determinar si la solución ofrece o no la funcionalidad sin entrar a valorar la eficiencia de las técnicas que cada solución aplica para proporcionar la funcionalidad.

Análisis de opiniones

Los proveedores de las soluciones de análisis de opiniones y escucha social no suelen disponer de información abierta y detallada respecto a las funcionalidades que ofrecen sus productos. Tras inspeccionar la documentación de las soluciones, contactar con los proveedores e incluso buscar noticias sobre anuncios

de nuevas funcionalidades, solo se ha conseguido analizar las funcionalidades de 7 de las 10 soluciones seleccionadas.

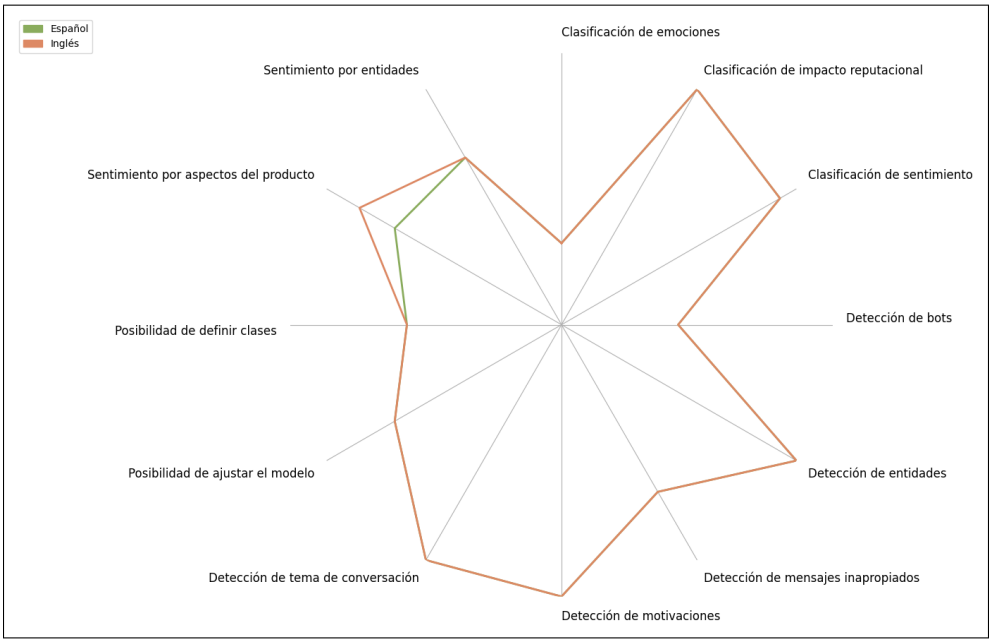

Figura 13: Comparativa de funcionalidades de soluciones de análisis de opiniones en inglés y español.

Tal y como puede observarse en la Figura 13, y al igual que en el Año 1, solo se ha identificado una brecha entre el inglés y el español en una de las soluciones donde la funcionalidad de sentimiento por aspecto de producto solo está disponible para textos en inglés. En el resto de las herramientas las funcionalidades en ambos idiomas son idénticas.

Las funcionalidades que más presencia tienen en las herramientas de análisis de opiniones son la clasificación de sentimiento, la detección de entidades, la detección de motivaciones y la detección de temas de conversación. Por el contrario, muy pocas ofrecen la funcionalidad de clasificación de tipos de emociones o la funcionalidad que permite definir clases de sentimientos. La Figura 13 ha sido generada a partir de los datos contenidos en la Tabla 26 del Apéndice D.

Asistentes virtuales

Los asistentes virtuales pueden ser muy diversos, comprendiendo modelos conversacionales como ChatGPT, asistentes virtuales tradicionales como Alexa o herramientas adaptables para diversas actividades como IBM Watson Asistant. Por lo tanto, resulta complicado realizar un mapa de funcionalidades homogéneo. Se han recogido las funcionalidades más relevantes: posibilidad de añadir habilidades, reconocimiento de voz, acentos o variantes regionales, capacidad de escribir texto y número de comandos aceptados. Las funcionalidades específicas de cada uno de los asistentes se han agrupado en *Otras funcionalidades*.

Como se puede observar en la Figura 14, la principal brecha se da en la oferta de variantes regionales que ofrecen los asistentes virtuales en inglés y en español. La mayoría de los asistentes virtuales pueden reconocer y entender diferentes acentos regionales de un idioma determinado, lo que facilita la comunicación para usuarios de diferentes regiones geográficas. Tanto el español como el inglés se hablan en una gran cantidad de países[35] y existen diferencias considerables entre las regiones, por ejemplo, entre el inglés de Estados Unidos, Reino Unido o Indio, o entre el español de México, Argentina o España. Los asistentes virtuales que ofrecen variantes regionales tienen, de media, tres veces más variantes para el inglés que para el español. La Figura 14 ha sido generada a partir de los datos de la Tabla 27 del

[35] *List of official languages by country and territory*, Wikipedia. 27 de febrero de 2023. Accedido: 6 de febrero de 2024. https://en.wikipedia.org/wiki/List_of_official_languages_by_country_and_territory

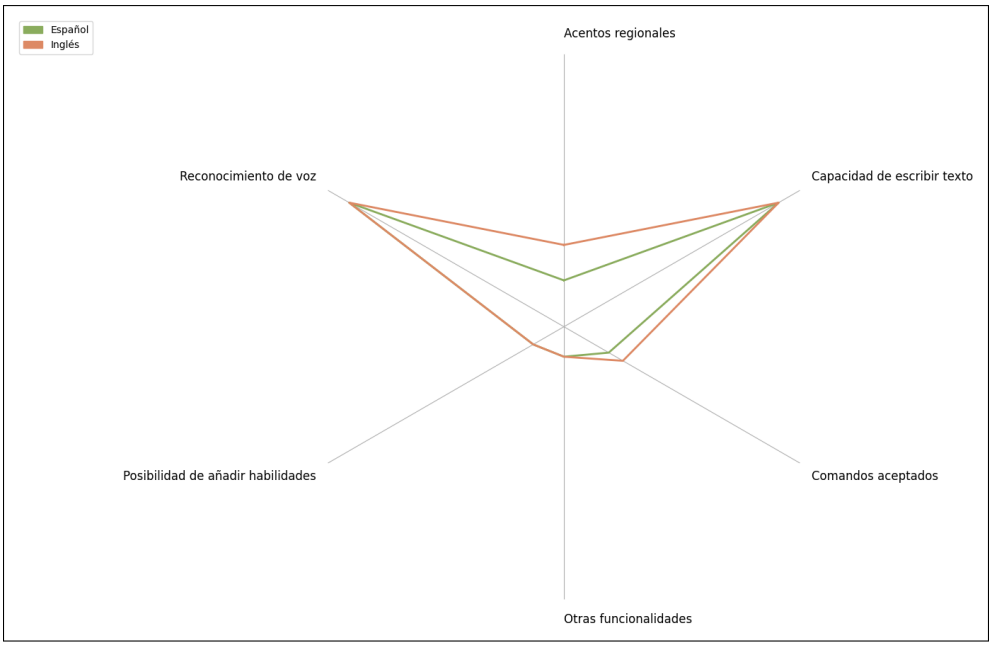

Figura 14: Comparativa de funcionalidades de asistentes virtuales en inglés y español.

Apéndice D.

Traducción automática

Como se puede ver en la Figura 15, no se han detectado diferencias relevantes en las funcionalidades de los traductores automáticos en inglés y español. Es común que los productos de inteligencia artificial usen traducción automática para ofrecer en más de un idioma funcionalidades que solo están disponibles en un idioma. Por ello, no resulta sorprendente que en las soluciones de traducción automática no se observen diferencias entre las funcionalidades ofrecidas en inglés y español. La única brecha se ha detectado en la funcionalidad de variantes regionales, ya que algunos traductores ofrecen más variantes regionales en español que en inglés o viceversa, resultando apenas favorable para el español.

Las funcionalidades de las que más traductores disponen son la opción de traducir de texto a texto y la opción de traducir documentos, como así también detectar idiomas. Por el contrario, las funcionalidades menos comunes son la traducción de voz a texto y de imágenes de palabras a texto, como así también la adaptación a dominios. La Figura 15 ha sido generada a partir de los datos contenidos en la Tabla 28 del Apéndice D.

Teclados predictivos

En el caso de los teclados predictivos 8 de las 9 soluciones analizadas tienen las mismas funcionalidades en ambos idiomas. La novena solución solo está disponible en inglés, de ahí la brecha observada en la Figura 16.

Las funcionalidades que todos los teclados ofrecen son la sugerencia de palabras, autocompletado de palabras y corrección gramatical. Algunos de los teclados sugieren expresiones o frases, sin embargo, solo uno de los analizados ofrece la funcionalidad en beta de sugerir textos de tamaño más grande como párrafos o secciones. La Figura 16 ha sido generada a partir de los datos de la Tabla 29 del Apéndice D.

Buscadores web

Existe una gran diferencia en las funcionalidades de los buscadores web de uso ciudadano y los buscadores corporativos. En general, los buscadores corporativos tienen menos funcionalidades de inteligencia

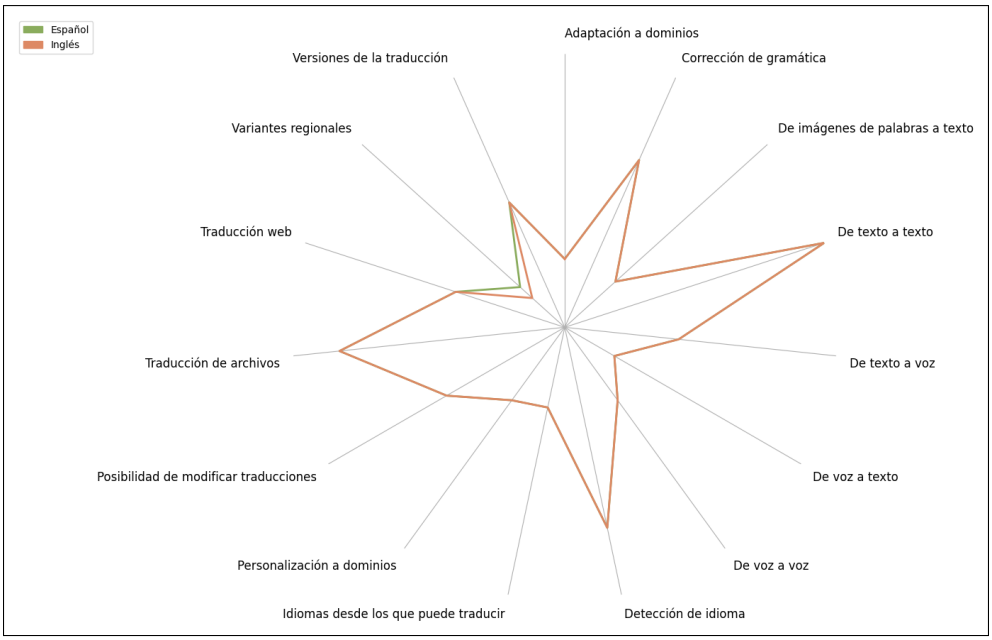

Figura 15: Comparativa de funcionalidades de soluciones de traducción automática en inglés y español.

artificial incorporadas, pero ofrecen la posibilidad de incluir modelos propios independientemente del idioma.

Tal y como se puede observar en la Figura 17, al igual que en el Año 1, la mayor brecha en las soluciones de búsqueda web se produce en la funcionalidad de búsqueda de sinónimos. También se mantuvo la diferencia en la corrección de gramática, mientras que se redujo la existente en clasificación de temas y búsqueda de respuestas.

La funcionalidad que más se repite en los buscadores es la detección de significado del texto. Mientras que en el lado opuesto, a diferencia del Año 1, se encuentra la funcionalidad de clasificación de temas ya que con los nuevos modelos de Inteligencia Artificial las soluciones han podido incorporar la búsqueda de texto en imágenes. La Figura 17 ha sido generada a partir de los datos de la Tabla 30 del Apéndice D.

6.3. Indicador S.1 Brecha en funcionalidades [S1: 8 %]

En base a las brechas identificadas en las funcionalidades de cada área de aplicación se calculan las brechas que se muestran en la Figura 18, y cuya media da una brecha de funcionalidades global de las soluciones de mercado analizadas del 8 %. Los detalles sobre la metodología para el cálculo del indicador I.S.1 se pueden encontrar en el documento “Ámbito 0.2 Diseño y cálculo de la métrica agregada para medir la brecha español/inglés en tecnologías de la lengua. Informe Año 2”.

Las soluciones con mayor cuota de mercado continúan disponibles tanto en inglés como en español, con la excepción de uno de los teclados predictivos que sigue sin ofrecerse para el español. A pesar de ello, al igual que en el Año 1, se sigue observando una brecha a nivel de funcionalidades que se distribuye de manera desigual entre las distintas áreas: muy reducida en los traductores automáticos y herramientas de análisis de opiniones; moderada en los buscadores web y teclados predictivos; y grande en los asistente virtuales. Sin embargo, a diferencia del Año 1, aunque poco, se redujo la brecha en buscadores web, teclados predictivos y asistentes virtuales a favor del español. Finalmente el indicador I.S.1 Brecha en funcionalidades es del el 8 %. Esto se da por la incorporación de nuevos modelos de lenguajes multilingue en varias de las herramientas en cuestión.

En relación al año anterior, la brecha es muy similar, pasando del 9 % al 8 %. Además, la diferencias de brecha entre áreas de aplicación es muy similar, siendo ésta muy reducida en los traductores automáticos

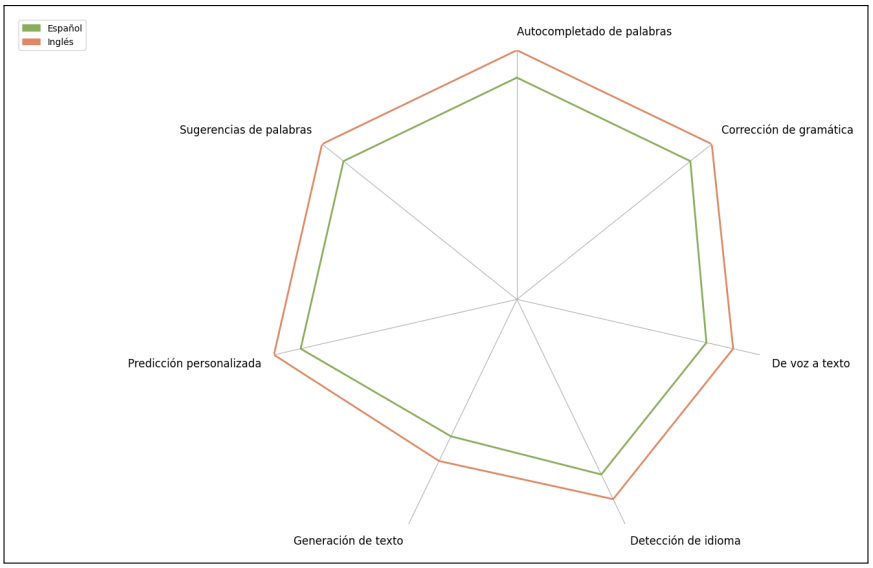

Figura 16: Comparativa de funcionalidades de teclados predictivos en inglés y español.

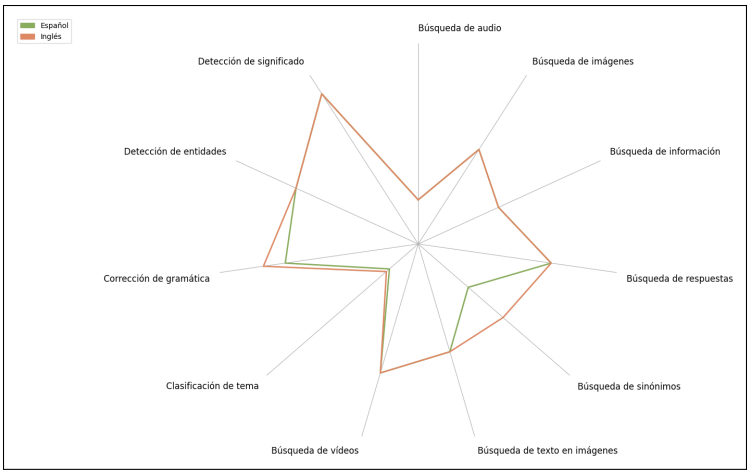

Figura 17: Comparativa de funcionalidades de buscadores web en inglés y español.

y herramientas de análisis de opiniones; moderada en los buscadores web y teclados predictivos; y grande en los asistente virtuales.

7. Cálculo de indicadores: Ámbito 3 - Nivel de Adopción

En esta sección se presentan los resultados relativos al nivel de adopción de tecnologías de la lengua por parte de empresas y de ciudadanos. Nos centramos en dar respuesta a dos preguntas fundamentales: i) ¿Existen diferencias en la adopción de tecnologías de la lengua por parte de empresas en función de la lengua que predomina en la empresa, inglés o español? ii) ¿Existen diferencias en la adopción de tecnologías de la lengua por parte de ciudadanos en función de la lengua principal que estos hablan?

Para contestar estas preguntas, se ha realizado un análisis de la adopción de herramientas de tecnologías del lenguaje en empresas de habla hispana e inglesa, utilizando una variedad de fuentes. En primer lugar, se han analizado menciones de herramientas de IA en presentaciones e informes de resultados corporativos de empresas. A continuación, se ha realizado un análisis similar de las menciones en medios de comunicación y se ha calculado el impacto de la adopción de tecnologías del lenguaje. Además, se han

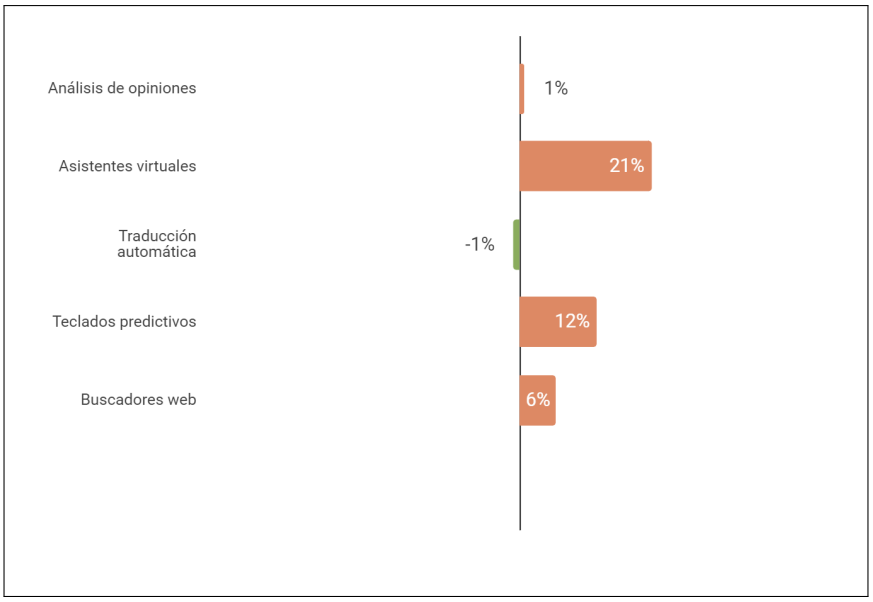

Figura 18: Brecha en funcionalidades por área de aplicación.

llevado a cabo encuestas en España y Estados Unidos para conocer el grado de adopción de herramientas por parte de los ciudadanos. En esta sección se presentan las conclusiones de este estudio, y el cálculo la brecha existente en el ámbito de adopción.

Las empresas seleccionadas y la metodología seguida para desarrollar y aplicar los indicadores se han descrito en la Sección 4.2.

7.1. Análisis de menciones en presentaciones e informes de resultados corporativos

En este apartado se analizan las menciones de las soluciones de mercado y de tecnologías del lenguaje en las presentaciones e informes de resultados corporativos de las empresas seleccionadas en el apartado anterior y publicados en los tres últimos años (2021, 2022 y 2023). Seleccionar los tres últimos años garantiza un volumen suficiente para que las brechas calculadas sean estadísticamente significativas.

7.1.1. Indicador A.1: Brecha en menciones de soluciones en informes corporativos [A1: 93 %]

Las áreas de aplicación y las soluciones analizadas para este estudio son aquellas establecidas en el documento “Ámbito 2 Soluciones de Mercado”.

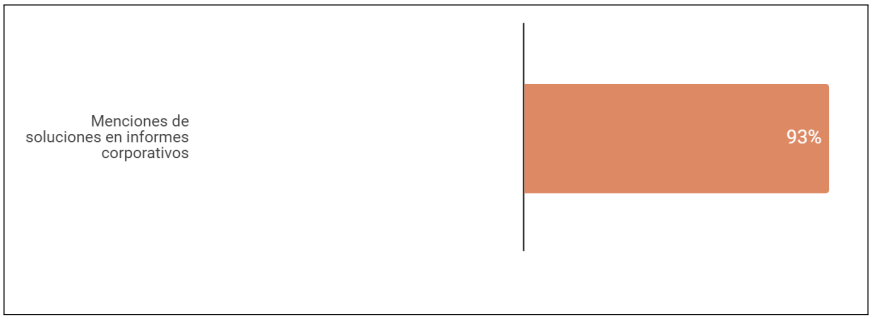

Figura 19: Brecha en menciones de soluciones en informes corporativos.

En la Figura 19 se muestra que la brecha en menciones de soluciones en informes corporativos es del 93 %. Cabe destacar que, de un año a otro, siguen siendo parte del análisis las mismas 19 soluciones desarrolladas por las cuatro grandes empresas que aparecen en la Tabla 1, las cuales se llevan gran parte

de las menciones: Apple, Microsoft, Alphabet y Amazon. Así mismo, solo se han encontrado 4 menciones a soluciones en las presentaciones e informes de resultados corporativos de las empresas españolas. Esto hace que la brecha se haya reducido muy poco respecto de la brecha del Año 1. El hecho de tener empresas que son desarrolladoras y adoptantes de las soluciones puede tener un impacto en la brecha, pero no hay volumen de coincidencias suficientes para determinarlo.

7.1.2. Indicador A.2: Brecha de menciones de tecnologías del lenguaje en informes corporativos [A2: 53 %]

A diferencia del Año 1, se han incrementado las menciones de tecnologías del lenguaje en los informes de empresas. Sin embargo, la brecha se redujo 3 puntos. Se han identificado 17 menciones en las empresas del índice S&P y 8 menciones en las del IBEX 35. En la Figura 20 se muestra que la brecha en menciones de tecnologías del lenguaje en informes corporativos es del 53 %.

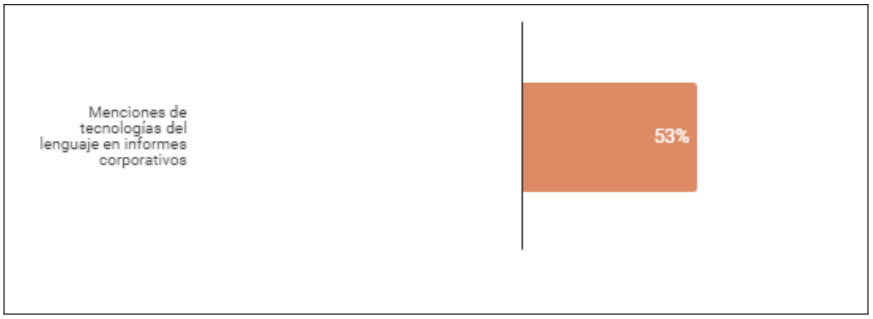

Figura 20: Brecha de menciones de tecnologías del lenguaje en informes corporativos.

7.2. Análisis de menciones en medios de comunicación

Adicionalmente, se han analizado las menciones en los medios de comunicación de las soluciones de mercado y de tecnologías del lenguaje en los medios de comunicación de las empresas seleccionadas en los tres últimos años (2021, 2022 y 2023).

7.2.1. Indicador A.3: Brecha en menciones de soluciones en medios de comunicación [A3: 83 %]

En total, las empresas del Índice S&P seleccionadas han sido mencionadas un total de 17.496.000 veces en los medios, mientras que las empresas del IBEX 35 seleccionadas han sido mencionadas un total de 3.276.000 veces en los medios. Para evitar sesgar las medidas por la popularidad de las empresas o la capacidad de generar noticias de los medios en un país u otro, las menciones han sido normalizadas dividiendo por el número total de noticias de las empresas de cada índice. Los datos usados para realizar el cálculo de brecha en las menciones de soluciones en medios de comunicación se encuentran el la Tabla 31 del Apéndice D.

En la Figura 21 se muestra la brecha en menciones de soluciones en medios de comunicación para cada una de las áreas de aplicación, cuya media da una brecha global del 83 % (Figura 22), habiendo subido un 7 % respecto del Año 1.

Al igual que en el análisis de menciones en informes corporativos, parte de la brecha corresponde a la adopción de herramientas de terceros y parte de la brecha corresponde a la adopción de herramientas propias. En este caso sí podemos detallar la brecha para cada una de las dos situaciones.

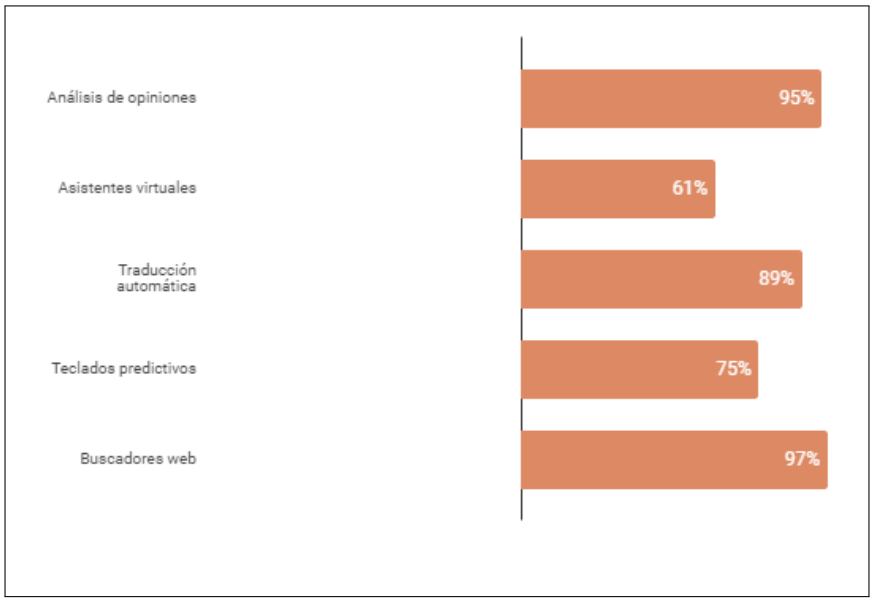

Figura 21: Brecha en menciones de soluciones en medios de comunicación por área de aplicación.

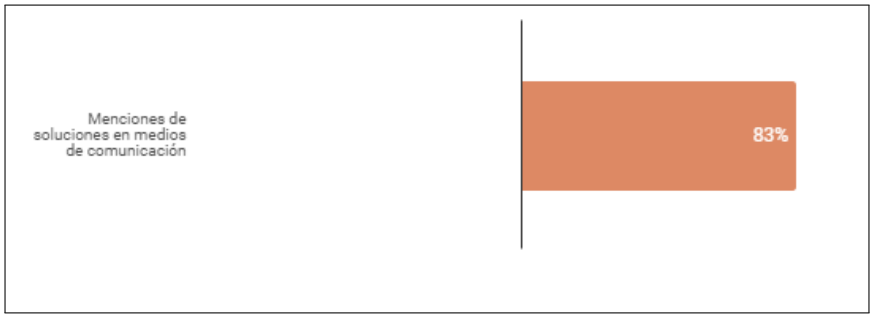

Figura 22: Brecha en menciones de soluciones en medios de comunicación.

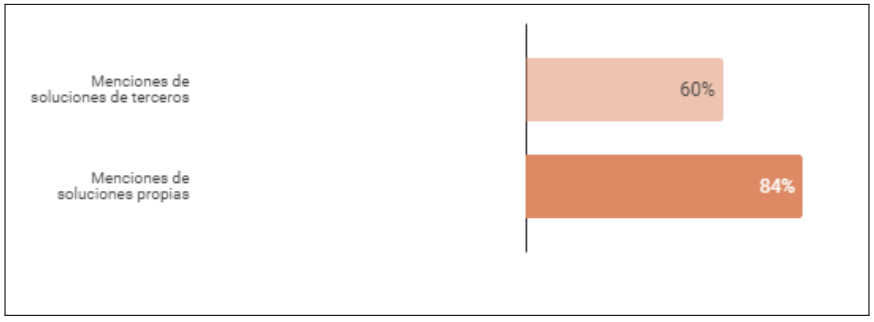

Figura 23: Brecha en menciones de soluciones en medios de comunicación por tipo de solución.

En la parte superior de la Figura 23, se muestra la brecha en menciones de herramientas de terceros, aquellas que no han sido desarrolladas por ninguna empresa de las presentes en la Tabla 1. En la parte inferior se muestra la brecha en menciones de las 19 herramientas que han sido desarrolladas por Apple, Microsoft, Alphabet o Amazon. Como se puede observar, a diferencia del Año 1 en ambos casos la brecha es del 60 %.

7.2.2. Indicador A.4: Brecha en menciones de tecnologías del lenguaje en medios de comunicación [A4: 70 %]

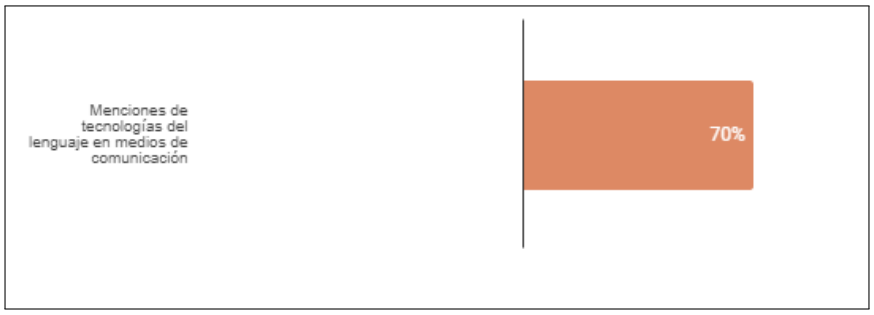

Figura 24: Brecha en menciones de tecnologías del lenguaje en medios de comunicación.

La brecha en menciones de tecnologías del lenguaje en medios de comunicación, que se muestra en la Figura 24, es del 70 %, habiendo incrementado en un 21 % respecto del Año 1. A su vez, es 13 puntos menor que la brecha de menciones de soluciones.

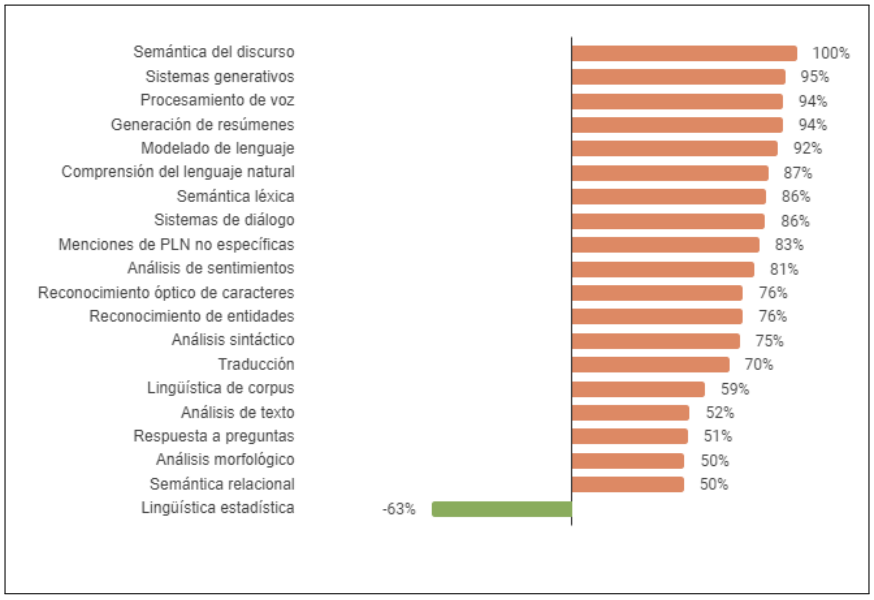

Figura 25: Brecha en menciones en medios de comunicación para cada una de las tecnologías del lenguaje analizadas.

Tal y como se puede ver en la Figura 25, en comparación con el Año 1, se mantiene únicamente la brecha a favor del español en lingüística estadística y en las tecnologías necesarias para crear soluciones como chatbots (respuesta a preguntas y sistemas de diálogo) se ha vuelto la brecha a favor del inglés. En el resto de las tecnologías, la brecha existente sigue siendo a favor del inglés y especialmente acentuada (mayor al 90 %) para: semántica de discurso, sistemas generativos, procesamiento de voz, generación de resúmenes y modelado de lenguaje.

7.2.3. Indicador A.5: Brecha en impacto de las tecnologías en la empresa [A.5: 60 %]

En la Figura 26 se muestra que la brecha en impacto de las tecnologías en la empresa es del 60 %, seis puntos menor que la brecha del Año 1. A diferencia del Año 1, dado el incremento en la brecha en

menciones de tecnologías del lenguaje en medios de comunicación (Figura 24), esta brecha comparada con aquella resulta un 23 % menor.

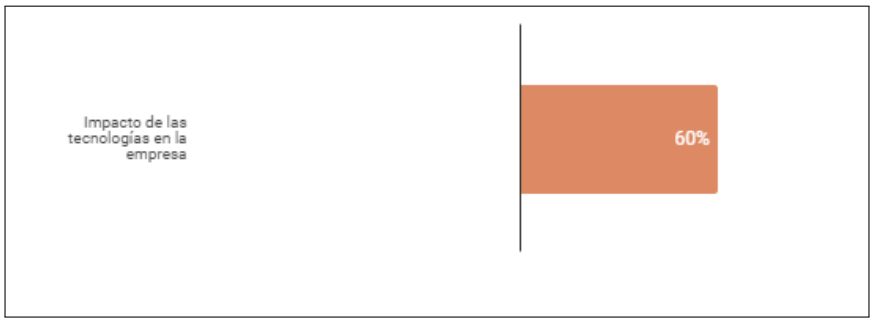

Figura 26: Brecha en impacto de las tecnologías en la empresa.

7.3. Medición del nivel de adopción basado en encuestas

Para medir el grado de adopción por parte de ciudadanos y empresas se han realizado del 8 al 12 de enero de 2024 1.805 encuestas sobre el uso de soluciones de tecnologías de la lengua en Estados Unidos y en España (904 y 901 respectivamente). El público objetivo fueron personas mayores de 18 años residentes en España y Estados Unidos para evaluar las soluciones en el idioma predominante de cada país.[36]

A cada una de las 1.805 personas encuestadas se le ha preguntado por todas las soluciones de las 5 áreas de aplicación seleccionadas. Se ha establecido la cifra de al menos 900 para cada idioma porque era el mínimo necesario para obtener resultados estadísticamente significativos.

Al igual que en el Año 1, las preguntas sobre el nivel de adopción de cada una de las áreas de aplicación se han dividido en dos bloques. En el primer bloque las preguntas se centran en el área de tecnologías de la lengua y dependen de si el encuestado ha utilizado alguna herramienta perteneciente al área, sin preguntar por una herramienta específica. En el segundo bloque las preguntas se centran en herramientas específicas y dependen de si el encuestado ha utilizado alguna de las herramientas específicas establecidas en documento “Ambito 2 Soluciones de Mercado Informe Año 2”. Las encuestas se adjuntan en el Apéndice C.

Los resultados de las preguntas del primer bloque se recogen en la Figura 27. En términos generales la adopción se mantiene estable respecto al Año 1 en ambos idiomas y para todas las herramientas. Cabe destacar que los buscadores web son el grupo de soluciones que tiene la mayor adopción en ambos idiomas, aunque en inglés es significativamente mayor que en español. Los teclados predictivos y los asistentes virtuales tienen un nivel similar de uso en ambos países. Las herramientas de traducción automática destacan significativamente en español, mientras que las de análisis de opiniones lo hacen en inglés.

Respecto a la adopción de herramientas por género, en base a los datos que se recogen en la Tabla 20, se observa en general poca diferencia en la adopción de herramientas entre hombres y mujeres. En inglés, sólo se observa una mayor adopción significativa de teclados predictivos en las mujeres que en los hombres. En la tabla no se presentan resultados del grupo que prefiere no identificarse con un género por lo reducido de su tamaño (menos del 1 % del total).

Respecto a la edad, como se puede observar en la Tabla 21, se observa el mismo patrón que en el Año 1 tanto en la adopción de herramientas para el español como el inglés. En ambos países e idiomas hay una clara tendencia a mayor adopción en los menores de 46 años para todas las herramientas excepto buscadores web.

En las siguientes secciones se analizarán los resultados de las preguntas pertenecientes al bloque 2, es decir, centradas en herrramientas específicas. Todos los datos con las respuestas de este segundo bloque se encuentran en la Tabla 32 del Apéndice D.

[36] En las encuestas se pregunta por el nivel de adopción, cuyos resultados se recogen en este apartado, y por la experiencia de usuario, cuyos resultados se recogen en el documento “Ambito 4 Experiencia de Usuario Informe Año 2”.

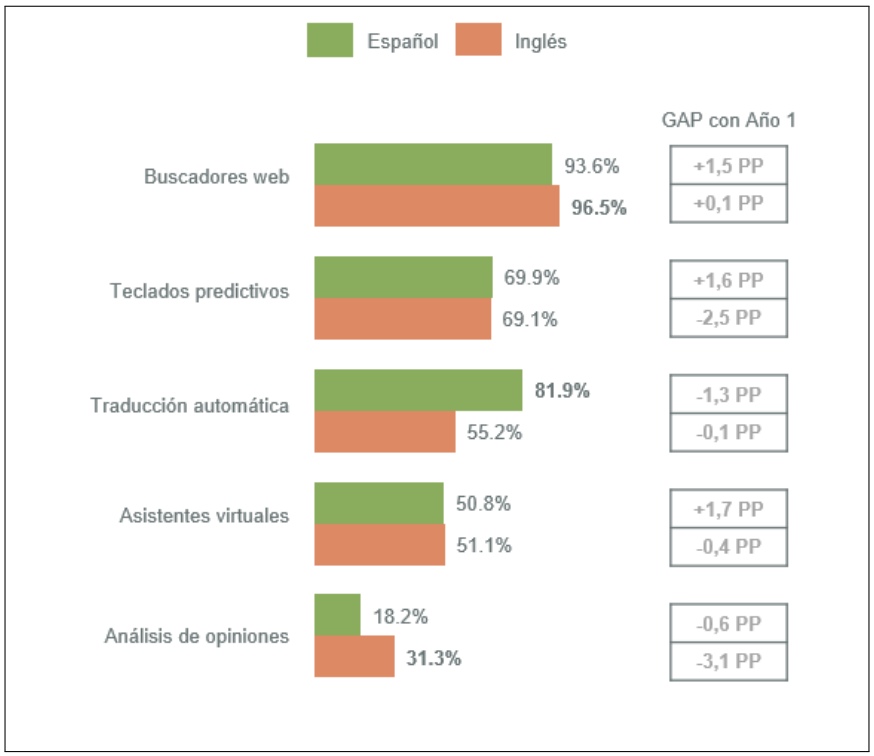

Figura 27: Resultados del primer bloque de las encuestas: adopción de cada una de las áreas de aplicación.

Tabla 20: Resultados del primer bloque de la encuesta: adopción de cada una de las áreas de aplicación por género.

	Español			Inglés		
	Total	**Masculino**	**Femenino**	**Total**	**Masculino**	**Femenino**
Análisis de opiniones	18.1 %	19.1 %	17.1 %	31.3 %	32.2 %	30.4 %
Asistentes virtuales	50.9 %	51.6 %	50.3 %	51.1 %	50.1 %	52.1 %
Traducción automática	82.0 %	82.9 %	81.1 %	55.2 %	54.0 %	56.4 %
Teclados predictivos	70.0 %	68.4 %	71.5 %	69.2 %	65.6 %	72.7 %
Buscadores web	93.5 %	92.4 %	94.7 %	96.5 %	97.2 %	95.7 %
Promedio	**62.9 %**	**62.9 %**	**62.9 %**	**60.7 %**	**59.8 %**	**61.5 %**

Tabla 21: Resultados del primer bloque de las encuestas: adopción de cada una de las áreas de aplicación por rango de edad.

	Español				Inglés			
	Total	**18 a 29**	**30 a 45**	**46 o más**	**Total**	**18 a 29**	**30 a 45**	**46 o más**
Análisis de opiniones	18.2 %	23.4 %	17.2 %	14.0 %	31.3 %	44.0 %	35.4 %	14.3 %
Asistentes virtuales	50.8 %	60.5 %	53.5 %	38.3 %	51.1 %	58.6 %	58.3 %	36.3 %
Traducción automática	81.9 %	86.8 %	82.2 %	76.7 %	55.2 %	69.2 %	60.3 %	36.0 %
Teclados predictivos	69.9 %	79.9 %	69.7 %	60.0 %	69.1 %	79.5 %	76.2 %	51.7 %
Buscadores web	93.6 %	93.8 %	97.0 %	90.0 %	96.5 %	94.7 %	96.7 %	98.0 %
Promedio	**62.9 %**	**68.9 %**	**63.9 %**	**55.8 %**	**60.6 %**	**69.2 %**	**65.4 %**	**47.3 %**

7.3.1. Análisis de opiniones

Dentro de la baja adopción general de las herramientas para nálisis de opiniones en ambos idiomas los niveles de adopción de estas herramientas presentan diferencias menores que no llegan a ser significativas en ningún caso.

En cuanto al español, el uso personal supera al profesional para Talkwalker y Khoros. El uso personal y profesional obtiene valores similares a los del uso personal para la mayoría de las herramientas. Mientras que en inglés, todas las herramientas muestran más uso personal que profesional. Khoros destaca en el uso conjunto (personal y profesional).

7.3.2. Asistentes virtuales

En asistentes virtuales, destaca la adopción de Alexa en ambos idiomas, manteniendo el primer lugar de un año a otro. ChapGPT presenta incrementos pronunciados y estadísticamente significativos en ambos idiomas.

En español, el uso personal supera de manera muy importante y en todas las herramientas al uso profesional. ChatGPT destaca en la adopción para uso conjunto. En cuanto a inglés, el uso personal supera de manera importante y en todas las herramientas al uso profesional.

7.3.3. Traducción automática

En traducción automática, Google Translate se mantiene en primer lugar en ambos idiomas, tal como en el Año 1. ChatGPT, que se mide por primera vez este año se posiciona en segundo lugar en ambos idiomas.

En español, se observa una tendencia a mayor uso personal que profesional o conjunto, especialmente en las herramientas de mayor adopción. En inglés, el uso personal supera de manera importante al uso profesional y a la proporción de personas que adoptan las soluciones para ambos usos.

7.4. Teclados predictivos

Gmail se mantiene en la primera posición en ambos idiomas, con variaciones leves respecto al Año 1. Microsoft Outlook decrece en español, mientras que SwiftKey decrece en inglés de manera significativa.

Se observa en español una tendencia a mayor uso personal frente al uso profesional y al uso conjunto. Las herramientas de Microsoft (Outlook y Office) son las únicas que superan ligeramente el uso personal con el uso conjunto. En inglés también destaca el uso personal frente al uso profesional y al combinado, a excepción de las herramientas de Microsoft (Outlook y Office) donde el uso profesional y conjunto es similar al uso personal.

7.4.1. Buscadores web

En buscadores web, Google Search mantiene la primera posición en ambos idiomas respecto del Año 1, en proporciones similares. Se observan diferencias muy leves entre el Año 1 y el actual en todas las herramientas y ambos idiomas.

El español, Google Search destaca con más de la mitad de sus usuarios reportando uso conjunto. Bing y Yahoo Search presentan sus mayores porcentajes de adopción para el uso personal. En inglés, el uso personal destaca frente al uso profesional y al conjunto. Google Search es el buscador con mayor proporción de uso conjunto.

7.5. Cálculo del Indicador A.6 Brecha en adopción de soluciones para uso profesional [A.6: 33 %]

Tal y como se puede observar en la Figura 28, existe una brecha entre el inglés y el español en el uso profesional de herramientas de análisis de opiniones del 70 % y en el uso de asistentes virtuales del 39 %. Esto implica una reducción de la brecha respecto de la del Año 1, del 10 % y 24 % respectivamente.

También presentan una brecha significativa a favor del inglés del 40 % en teclados predictivos y 18 % en buscadores web, un 4 % y 8 % menos respecto que la brecha del año anterior. En cuanto a los traductores automáticos se redujo la brecha y aunque es mínima, se volvió a favor del español. La media da una brecha en adopción de soluciones para uso profesional del 33 %, un 13 % menos que la del año pasado.

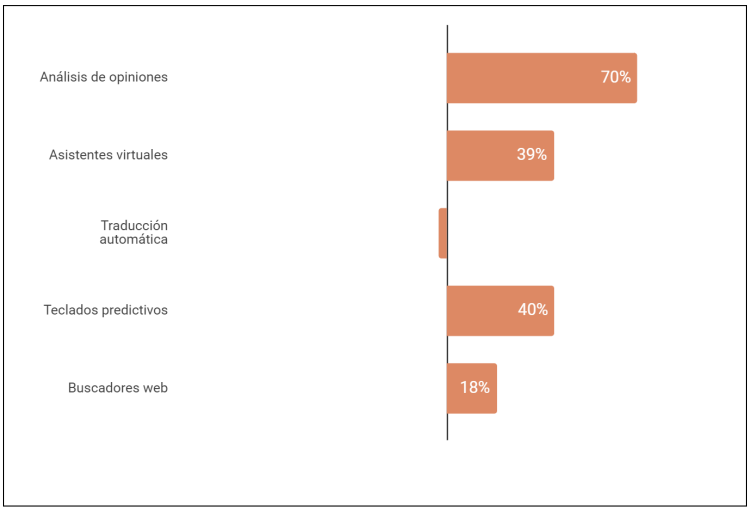

Figura 28: Resultados del segundo bloque de las encuestas: brecha en adopción de las soluciones de cada una de las áreas de aplicación para uso profesional.

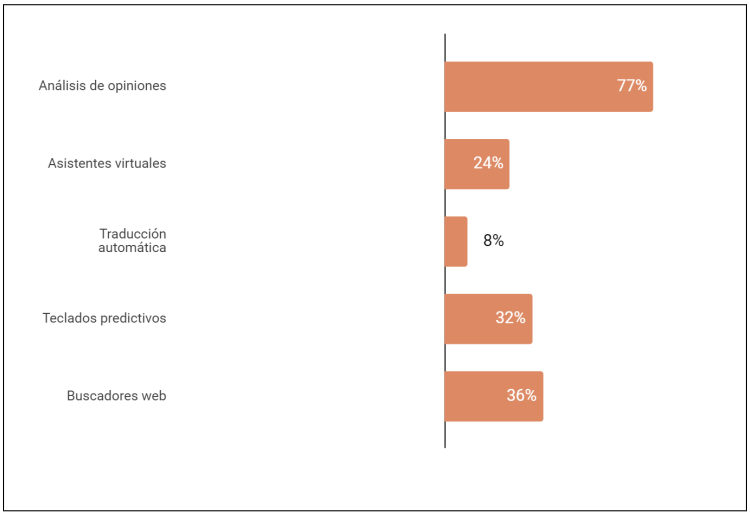

Figura 29: Resultados del segundo bloque de las encuestas: brecha en adopción de las soluciones de cada una de las áreas de aplicación para uso personal.

7.6. Cálculo del Indicador A.7 Brecha en adopción de soluciones para uso personal [A7: 36 %]

Como se puede observar en la Figura 29, no existe brecha de traductores automáticos, mientras que en el resto de tipos de herramientas sí que existe y es a favor del inglés. Hay una brecha significativa a favor del inglés de un 77 % para análisis de opiniones, de un 25 % para asistentes virtuales, un 28 % para teclados predictivos y de un 32 % para bsucadores web. La media de cada una de las áreas de aplicación da una brecha en adopción de soluciones para uso personal del 36 %, un 3 % mayor que en el Año 1.

7.7. Conclusiones y evolución de la brecha

En base a los resultados obtenidos con la metodología aplicada para este estudio, se concluye que sigue habiendo una brecha alta en la adopción de soluciones y tecnologías de IA en español respecto al inglés. La brecha es en promedio del 75 % a favor de este último idioma con un impacto del 60 %.

Por otra parte, al igual que en el Año 1, los ciudadanos de habla inglesa manifiestan hacer un mayor uso de soluciones de IA basadas en tecnologías del lenguaje que los ciudadanos de habla hispana. Esta brecha a diferencia del año anterior, es mayor en el uso personal (por solo un 3 %) ya que la brecha del uso profesional se redujo un 13 %.

Si observamos todos los indicadores de adopción, aunque en algunos sí que se redujo levemente la brecha, no ha alcanzado para reducir la brecha general de soluciones y tecnología de IA tanto para uso personal como el profesional en español respecto al inglés. La media de los indicadores A.1, A.2, A.3, A.4, A.5, A.6 y A.7, nos da una brecha global de nivel de adopción del 61 %, un 1 % más que la brecha general calculada en el año anterior.

8. Cálculo de indicadores: Ámbito 4 - Experiencia de Usuario

En esta sección se presentan los resultados del análisis de la brecha en experiencia de usuario. La experiencia de usuario se mide por medio de cuatro indicadores, dos de los cuales son calculados a partir de las opiniones y reseñas que los usuarios dejan de las soluciones que éstos utilizan y otros dos indicadores que son calculados a partir de las respuestas que se obtienen de cuestionarios que se realizan a usuarios de las soluciones en cuestión.

La Sección 8.1 se centra en medir las diferencias en la polaridad de opiniones entre los usuarios de tecnologías de la lengua en inglés y español. En la Sección 8.2 se miden los principales atributos mencionados por los usuarios en cada una de las áreas de aplicación. En la Sección 8.3 se muestran los detalles de satisfacción de usuario y las limitaciones de uso encontradas en las soluciones de tecnologías de la lengua seleccionadas.

8.1. Análisis de opiniones y reseñas

En esta sección se realiza una descripción de los datos obtenidos y se calcula el indicador de brecha en polaridad reputacional.

8.1.1. Obtención de opiniones

Para la obtención de los **mensajes de redes sociales** se ha utilizado la herramienta Brandwatch[37] que permite la extracción de datos mediante consultas. En las herramientas donde el uso principal se corresponde con la funcionalidad que se quiere estudiar, por ejemplo, el caso de la funcionalidad de traducción de DeepL, la consulta se ha realizado con el propósito de obtener todas las opiniones sobre dicha marca. En las herramientas cuyo uso principal no es el de la funcionalidad que se quiere estudiar, por ejemplo, el caso de la funcionalidad de escritura predictiva de GMail, Outlook, Google Workspace o Microsoft Word, la consulta se ha realizado con el propósito de obtener los mensajes que mencionan tanto la marca como la funcionalidad, sacrificando cobertura para tener mayor precisión.En todos los casos en que el nombre de la solución pudo resultar ambiguo, como lo son el de Alexa o Resonate, se han tomado medidas para desambiguar su nombre acompañándolo por otros términos o expresiones utilizadas para

[37] https://www.brandwatch.com/

hacer referencia a la solución. En el caso de Alexa lo son por ejemplo el término "amazon" o "asistente virtual" ("virtual assistant" en inglés).

Para la obtención de las reseñas, en primer lugar, se han identificado aquellas soluciones que tienen aplicaciones móviles en Google Play[38] o App Store.[39] En segundo lugar, se ha buscado si la solución tiene una página de reseñas en G2.[40] Por último, se han utilizado *scrapers* desarrollados por el área de Deep Learning[41] de LLYC[42] para extraer las reseñas de todas las aplicaciones móviles identificadas y un *scraper* de Apify[43] para las páginas de reseñas de G2 encontradas.

En total se han analizado 576.463 opiniones del año 2023 de las fuentes seleccionadas. El detalle para cada uno de los idiomas y fuentes puede encontrarse en la Tabla 22.

Tabla 22: Número de mensajes y reseñas analizados por idioma.

Fuente	Español	Inglés
Twitter	49.470	109.597
Reddit	3.715	56.567
Tumblr	1.651	39.621
Blogs	16.588	42.350
Foros	9.942	56.016
Google Play	74064	82.249
App Store	2.201	28.370
G2	406	3.656
Total	**158.037**	**418.426**

Para el análisis se han seleccionado aquellas soluciones de las que se han logrado obtener, al menos, 100 opiniones en cada idioma. La lista final de las aplicaciones seleccionadas abarca herramientas de todas las áreas de aplicación y es la siguiente:

- Análisis de opiniones: Brandwatch, Digimind, Meltwater, NetBase Quid, Sprinklr, Talkwalker.
- Asistentes virtuales: Alexa, Bixby, ChatGPT, Google Assistant, Google Bard, Siri.
- Traducción automática: Microsoft Translator o Bing Translator, DeepL, Google Translate, memoQ Translator PRO, Smartling, Reverso Translation.
- Teclados predictivos: Fleksy, GBoard, GMail, iPhone Keyboard, Microsoft Office 365, Microsoft Swiftkey, Microsoft Outlook, Grammarly.
- Buscadores web: Bing, Brave Search, DuckDuckGo, Elasticsearch, Google Search, Perplexity, Yahoo Search.

8.1.2. Indicador E.1: Brecha en polaridad reputacional [E.1: -9 %]

Como muestra la Figura 30 la brecha es a favor del español en todas las áreas de aplicación, salvo en la de traducción automática que a pesar de resultar la brecha a favor del inglés no es significativa. Cabe destacar que la brecha a favor del español en asistentes virtuales, aumentó un 17 % respecto del Año 1. En el resto de áreas el aumento a favor del español es menor: un 8 % en análisis de opiniones y buscadores web, mientras que en teclados predictivos solo un 1 %. Es importante recordar que la brecha sea de -19 % indica que la relación de opiniones positivas y negativas se decanta más hacia las positivas

[38] https://play.google.com/store/apps
[39] https://www.apple.com/app-store/
[40] https://www.g2.com/
[41] https://llyc.global/en/capability/deep-learning/
[42] https://llyc.global/
[43] https://apify.com/

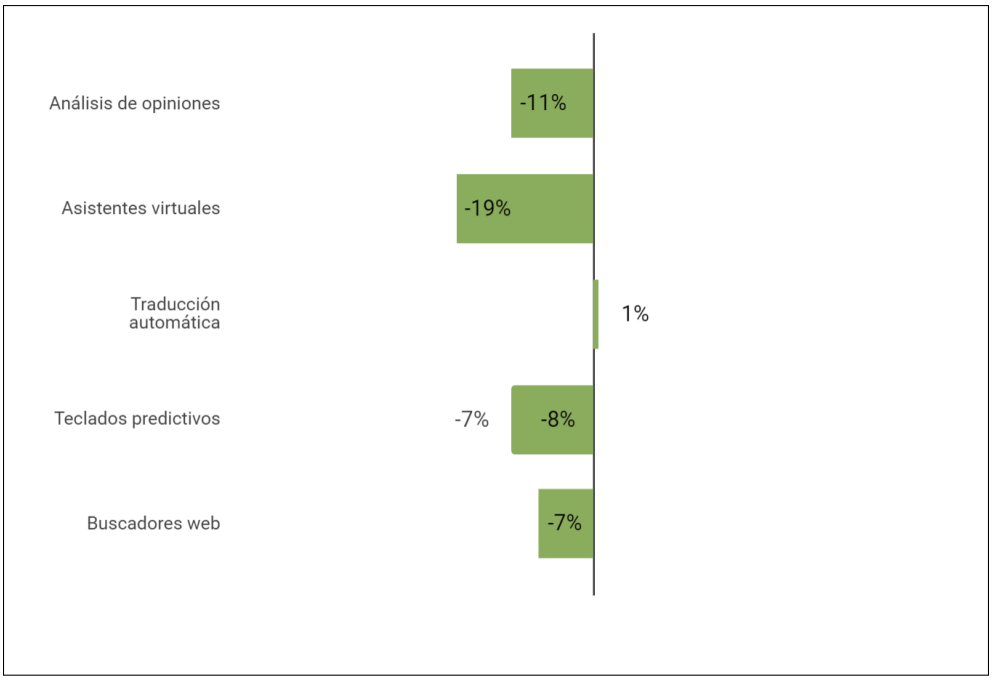

Figura 30: Brecha en polaridad reputacional por área de aplicación.

en las opiniones en español que en las opiniones en inglés. Finalmente, tal como muestra la Figura 31, la brecha en polaridad reputacional global de las soluciones analizadas es del -9 %, un 7 % más a favor del español que en el Año 1. Todos los datos para generar la Figura 30 se encuentran en la Tabla 33 del Apéndice D.

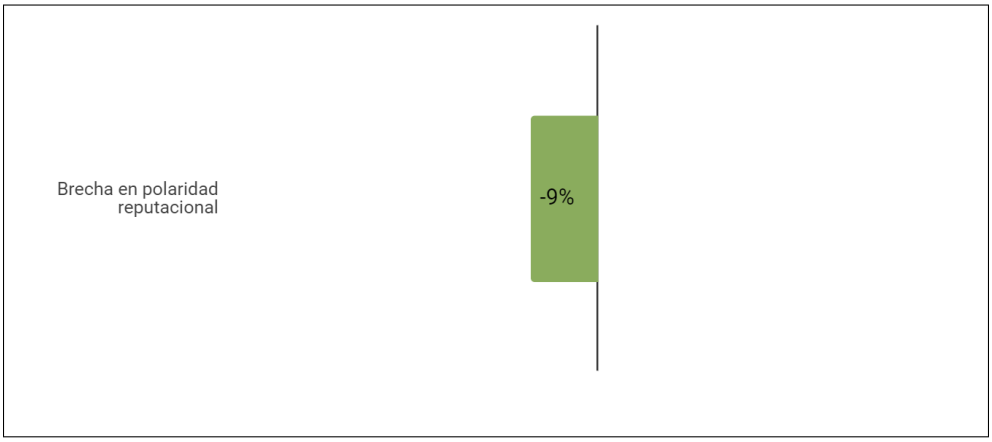

Figura 31: Brecha en polaridad reputacional.

8.2. Análisis de las curvas de valor

En este apartado, en primer lugar, se realiza un análisis de las curvas de valor de cada una de las áreas de aplicación. Después, se realiza el análisis de las curvas de valor globales. Por último, se calcula la brecha en curvas de valor. Se analizan cuatro atributos del producto en base a la ocurrencia de ciertos términos clave que se han identificado mediante expresiones regulares que se encuentran en el Apéndice B. El número de mensajes y reseñas en los que se ha detectado cada uno de los atributos, con los que se han calculado dichas curvas se encuentran en la Tabla 23.

Tabla 23: Número de mensajes por área de aplicación, atributo identificado e idioma.

Área	Atributo	Español	Inglés
Análisis de opiniones	Rendimiento	199	1.292
	Usabilidad	93	510
	Seguridad y privacidad	405	1.011
	Precio	68	719
Asistentes Virtuales	Rendimiento	3.466	6.694
	Usabilidad	4.770	6.717
	Seguridad y privacidad	1.207	3.121
	Precio	1.690	1.893
Traducción Automática	Rendimiento	1.733	4.830
	Usabilidad	913	3.391
	Seguridad y privacidad	186	1.434
	Precio	357	677
Teclados Predictivos	Rendimiento	1.671	4.446
	Usabilidad	1.582	4.832
	Seguridad y privacidad	493	5.418
	Precio	712	1.390
Buscadores Web	Rendimiento	2.280	6.709
	Usabilidad	856	6.876
	Seguridad y privacidad	1.264	5.968
	Precio	560	1.695
Total	**Rendimiento**	**9.349**	**23.971**
	Usabilidad	**8.214**	**22.326**
	Seguridad y privacidad	**3.555**	**16.952**
	Precio	**3.387**	**6.374**

A continuación se analizarán las curvas de valor de cada una de las áreas de aplicación y las curvas de valor globales. Los datos con los que se han calculado dichas curvas se encuentran en la Tabla **??** del Apéndice D.

8.2.1. Análisis de opiniones

Al analizar la curva de valor de las soluciones de análisis de opiniones de la Figura 32 y comparándola con la obtenida en el Año 1, se observa que los usuarios hispanos siguen valorando mejor los atributos de rendimiento, usabilidad y seguridad/privacidad por sobre el atributo precio, el cual valoran peor. En cuanto al margen respecto a cómo los usuarios anglosajones valoran los atributos, el margen se redujo con especial énfasis en el atributo rendimiento. Por otro lado, que el mayor margen se encuentre en el atributo precio, se puede deber a que este tipo de herramientas sea principalmente de uso empresarial, tengan un alto coste comparado con el software de uso cotidiano y que el poder adquisitivo de los países de habla inglesa sea mayor que el de los países de habla hispana.[44]

8.2.2. Asistentes virtuales

Como se puede observar en la Figura 33 y a diferencia del Año 1, aumentaron las diferencias en las curvas de valor de los asistentes virtuales en inglés y en español para los atributos de rendimiento, usabilidad y seguridad/privacidad. Los países de habla inglesa siguen teniendo una peor percepción del atributo precio. Recordar que esta área tiene una combinación de herramientas de uso ciudadano de coste bajo y herramientas de uso empresarial de coste alto con mayor adopción en países de habla inglesa.

[44] https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita

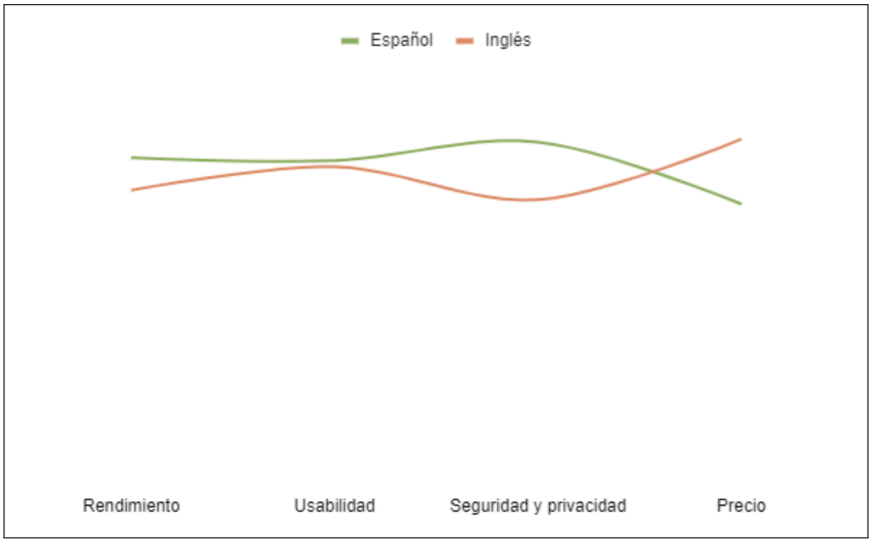

Figura 32: Curvas de valor de soluciones de análisis de opiniones.

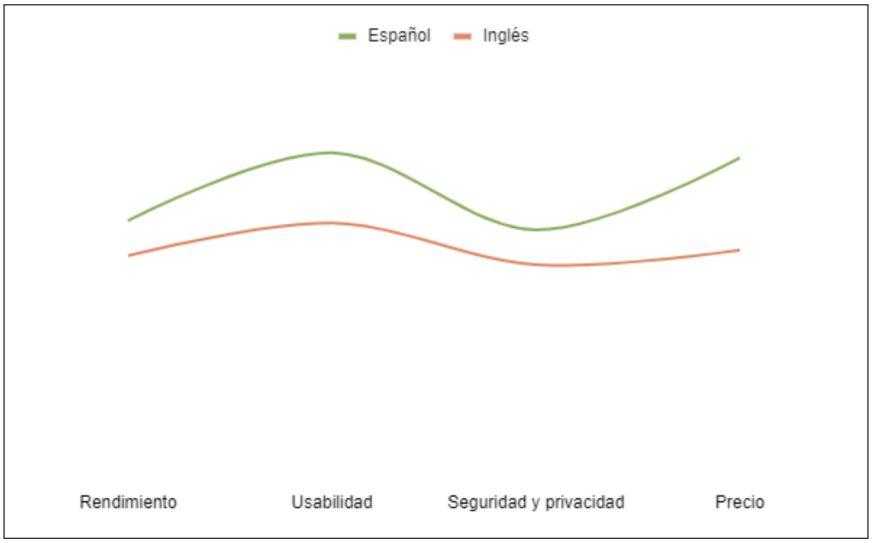

Figura 33: Curvas de valor de asistentes virtuales.

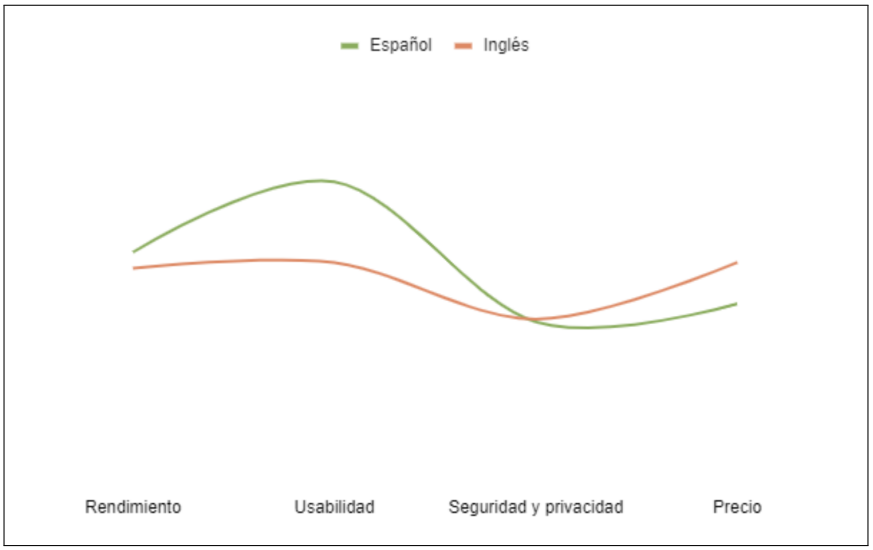

Figura 34: Curvas de valor de soluciones de traducción automática.

8.2.3. Traducción automática

Como se observa en la Figura 34 y al igual que en el Año 1, sigue sin haber diferencias relevantes en las curvas de valor de las soluciones de traducción automática en inglés y en español. El único punto a destacar es que los usuarios de habla hispana valoran mejor la usabilidad de las herramientas a pesar de que las funcionalidades, tal y como se muestra en el documento "Ámbito 2 Soluciones de mercado", son las mismas.

8.2.4. Teclados predictivos

Como se observa en la Figura 35, los países de habla hispana valoran mejor los atributos usabilidad y rendimiento mientras que valoran peor los de seguridad/privacidad y precio a diferencia de los de habla inglesa. Cabe destacar que al igual que en el Año 1, se mantienen la amplia brecha en el atributo de usabilidad a favor del español y la brecha significativa en el atributo seguridad/privacidad a favor del inglés.

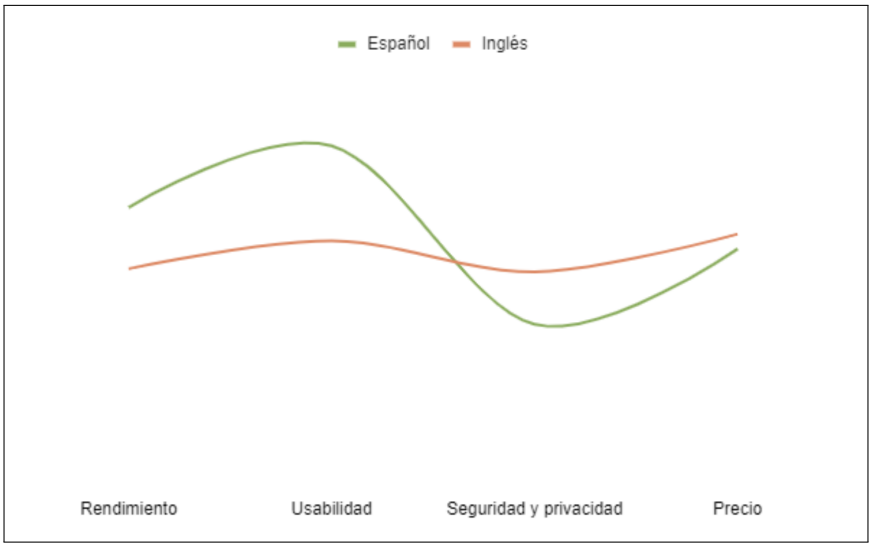

Figura 35: Curvas de valor de teclados predictivos.

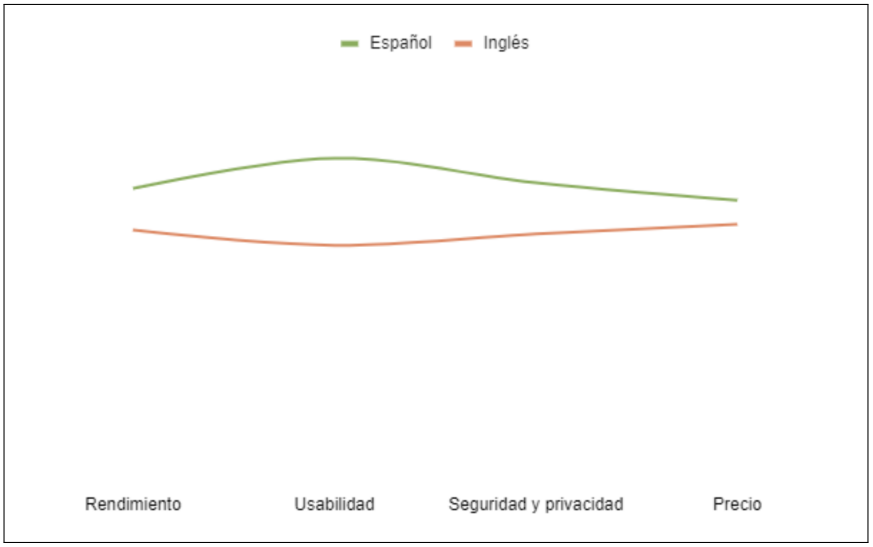

Figura 36: Curvas de valor de buscadores web.

8.2.5. Buscadores web

Como podemos observar en la Figura 36, y a diferencia del Año 1, en los países de habla hispana ya se valoran mejor los cuatro atributos, con mayor énfasis en el atributo usabilidad.

8.2.6. Curvas de valor globales

En las curvas de valor globales que se pueden ver en la Figura 37, se observa que al igual que en el Año 1, la mayor brecha se produce en la percepción de la usabilidad. En cambio la brecha para el rendimiento y la seguridad/privacidad es mucho más reducida y es nula para el precio. Si bien hubo algunas diferencias por área de aplicación de un año a otro, en términos generales la situación es similar.

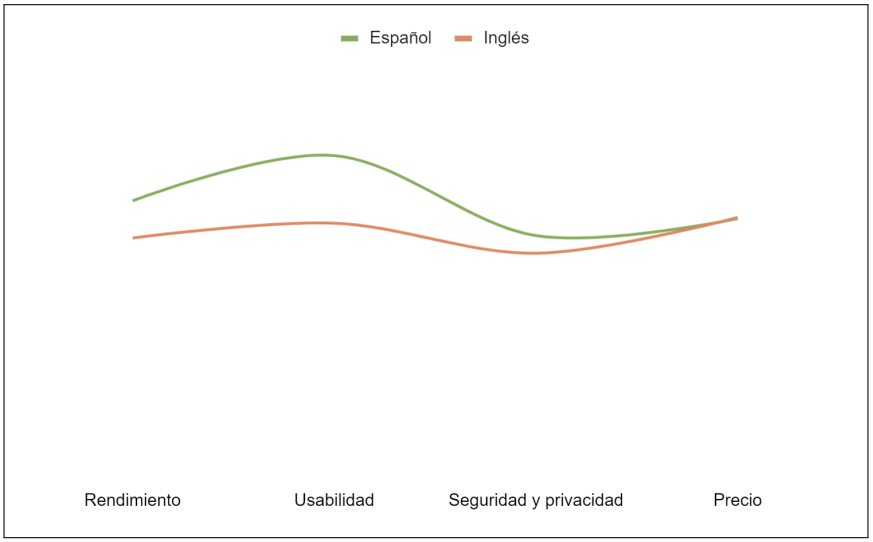

Figura 37: Curvas de valor globales.

Los puntos en los que podría mejorarse la propuesta de valor de las soluciones en español siguen siendo los propuestos en el año 1 del proyecto:

1. **Precio:** Tal y como se ha indicado el precio es el aspecto que los usuarios de habla hispana peor valoran en aquellas áreas en las que hay herramientas corporativas de alto coste. Esto puede deberse a la diferencia de poder adquisitivo medio entre países hispanos y anglosajones . El precio es el aspecto mejor valorado por los usuarios de habla inglesa con un margen del 7 % respecto a usabilidad; pero es el segundo, con un margen del -9 % respecto a usabilidad, para los usuarios de habla hispana.

2. **Seguridad y privacidad:** Es el aspecto peor valorado por usuarios de herramientas en ambos idiomas Es el aspecto peor valorado por usuarios de herramientas en ambos idiomas.

3. **Rendimiento:** Para los usuarios de habla hispana, el rendimiento de las soluciones tiene una valoración mucho más baja, del -17 %, que usabilidad. La diferencia es menos pronunciada, del -7 %, con los usuarios de habla inglesa.

8.2.7. Indicador E.2: Brecha en curvas de valor [E.2: 9 %]

En la Figura 38 se representan las brechas en curvas de valor calculadas para cada una de las áreas de aplicación analizadas anteriormente. Existen brechas a favor del español en todas las áreas de aplicación.

A diferencia del Año 1, en el cual la mayor brecha se dio en las soluciones de análisis de opinión, este año se ha reducido de un -15 % (a favor del español) a un -2 %. La mayor brecha se observa este año en soluciones de asistentes virtuales a favor del español, siendo del -19 % y habiendo aumentado en un 16 %.

En las soluciones de traducción automática la brecha se redujo de -7 % a favor del español a un -3 %, mientras que en buscadores web la brecha aumentó a favor del español de -7 % a -16 %. En el caso de teclados predictivos, si bien la brecha es algo baja pero a favor del español, cabe destacar que en el Año 1 lo era mayor y a favor del inglés.

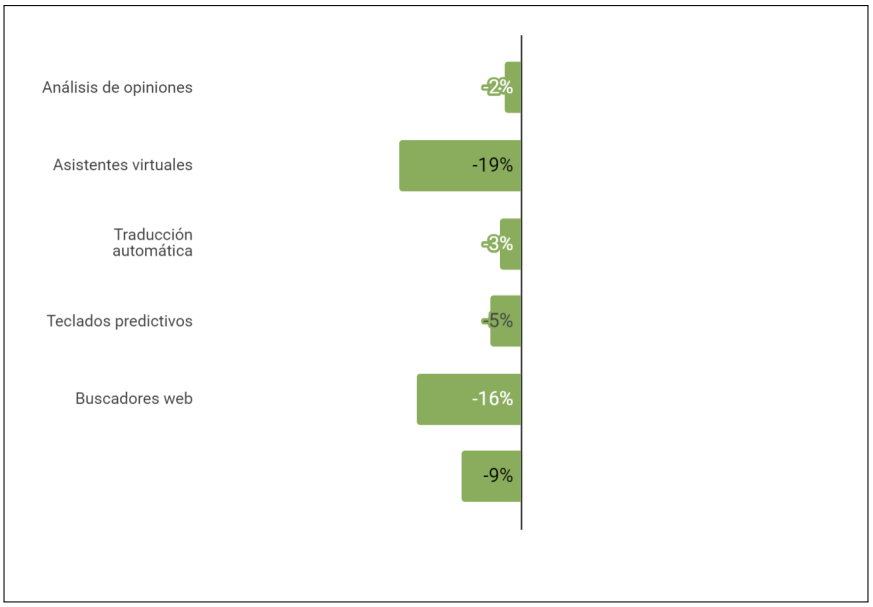

Figura 38: Brecha en curvas de valor por área de aplicación.

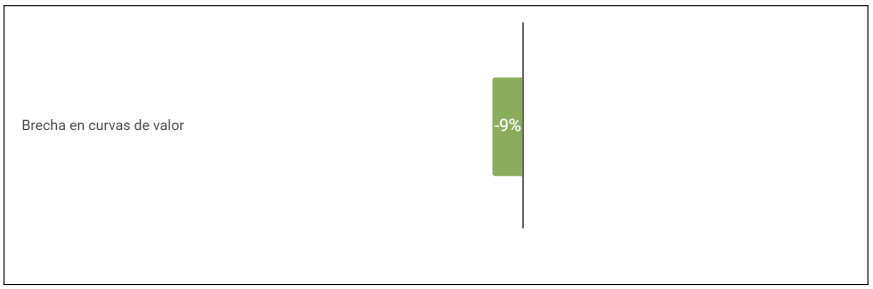

Figura 39: Brecha en curvas de valor.

Finalmente y como se indica en la Figura 39 la media da una brecha en curvas de valor global del -9 % lo que significa un 5 % más a favor del español que en el Año 1.

8.3. Satisfacción de usuario en encuestas

Para obtener datos sobre la satisfacción de usuario, se han realizado encuestas con preguntas sobre la experiencia de usuario y el uso de soluciones en Estados Unidos y en España, garantizando un mínimo de 900 encuestas en cada uno de los países. Los resultados se presentan en este apartado. Previamente en el documento "Ambito 3 Nivel de adopción Informe Año 2" se han presentado los resultados de las encuestas para el nivel de adopción.

A cada una de las 1800 personas encuestadas se le ha preguntado por todas las soluciones de las cinco áreas de aplicación. Se ha establecido la cifra de al menos 900 para cada idioma porque es el mínimo necesario para obtener resultados estadísticamente significativos.

Las preguntas sobre la satisfacción de usuario de cada una de las áreas de aplicación se han dividido en dos bloques:

1. Preguntas sobre el nivel de satisfacción con cada una de las herramientas que el usuario manifiesta haber usado en las preguntas referentes al "Ámbito 3 Nivel de adopción". Cada encuestado ha valorado la satisfacción de la herramienta con una puntuación del 1 al 5.

2. Preguntas sobre las limitaciones encontradas en cada una de las herramientas que el usuario manifiesta

haber usado en las preguntas referentes al "Ámbito 3 Nivel de adopción". Cada encuestado ha podido seleccionar las limitaciones que ha observado en el uso de la herramienta.

En las siguientes subsecciones se realiza un análisis de cada uno de los bloques y se calculan los indicadores correspondientes.

Todos los datos obtenidos de las encuestas sobre la satisfacción de usuario se encuentran en las Tablas 34 y 35 del Apéndice D.

8.3.1. Indicador E.3: Brecha en satisfacción de usuario [E.3: 12 %]

En cuanto a las soluciones de **análisis de opinión**, el nivel de satisfacción manifestada por los encuestados sobre el uso de las herramientas es mayor en Estados Unidos que en España en todos los casos. Hay que tener en cuenta que en este último país, la cantidad de personas que adopta este tipo de herramientas continúa siendo baja. En Estados Unidos, todas las soluciones presentan un nivel de satisfacción media por encima de 4 sobre 5, mientras que en España se sitúa algo por debajo de los 4 puntos sobre 5.

Los niveles de satisfacción con los **asistentes virtuales** continúan siendo ligeramente superiores entre los usuarios de Estados Unidos. Salvo Bixby, todos los asistentes en este último país están valorados por encima de 4 puntos sobre 5. En cambio en España, la valoración se sitúa entre 3 y 4 puntos sobre 5 en este tipo de aplicaciones. Cabe destacar que en España el asistente mejor valorado es ChatGPT mientras que en Estados Unidos el primer puesto lo comparten Amelia y Amazon Lex. En el caso de España el asistente mejor valorado, ChatGPT, coincide con el asistente que destaca en el uso combinado (para uso personal y profesional), mientras que en Estados Unidos no sucede lo mismo. Sin embargo, la valoración de la satisfacción de los asistentes más adoptados (Alexa, Siri y Google Assistant) es alta.

Entre las soluciones de **traducción automática**, en Español continúa destacando DeepL por su nivel de satisfacción, con una valoración media 4,19 sobre 5; seguido por ChatGPT con 3,85 y Google Translate con 3,76 sobre 5. En Estados Unidos sucede que varios traductores tienen una valoración de la satisfacción muy similar. Entre los mejores valorados en este país se encuentran Google Translate, DeepL, Amazon Translate, memoQ Translator y ChatGPT.

En Español, los **teclados predictivos** tienen valoraciones similares entre sí y respecto a las del Año 1: entre 3,5 y 4,0 sobre 5. En Estados Unidos observamos un rango ligeramente mayor: sobrepasando los 4 sobre 5.

En el área de **buscadores web**, Google destaca al igual que en el Año 1 en primer lugar con 4,26 y 4,43 sobre 5 en España y Estados Unidos respectivamente y supera con gran distancia al resto de los buscadores web valorados en España por debajo de 4 y en Estados Unidos según el buscador ya que los únicos que no superan los 4 puntos de satisfacción son Yahoo Search y Brave Search.

En la Figura 40 se muestra la brecha de satisfacción de usuario de cada una de las áreas de aplicación.

En conclusión, la brecha en la satisfacción de usuario si bien no es demasiado alta, es mayor a favor del inglés en todas las áreas. La media de todas las áreas de aplicación da una brecha en la satisfacción de usuario global del 12 %, un 10 % mayor que la brecha del Año 1. Todos los datos pueden encontrarse en la Tabla 34 del Apéndice D. El diseño de las encuestas realizadas puede encontrarse en el documento "Ámbito 3 Nivel de adopción Informe Año 2".

8.3.2. Indicador E.4: Brecha en limitaciones de uso

En **análisis de opiniones**, la mayoría de los usuarios percibe algún tipo de limitación en las herramientas que utiliza. Sin embargo, no se observa en general una tendencia en el tipo de limitaciones. Es importante destacar que el bajo número de usuarios en España hace poco robusta la muestra que dificulta la interpretación de las las diferencias entre España y Estados Unidos. Linkfluence es la única herramienta donde se presentan diferencias significativas entre las limitaciones reportadas en España y en USA. Para los encuestados en USA es mayor la limitación "Solo aporta información relativamente obvia" que para los de España. Todos los datos recogidos se encuentran en la Tabla 36 del Apéndice D.

En cuanto a los **asistentes virtuales**, en general, una parte importante de los usuarios percibe alguna limitación en las herramientas que usa. No se observa una tendencia en el tipo de limitaciones. Google Assistant destaca por ser la herramienta con menor percepción de limitaciones tanto en España como en

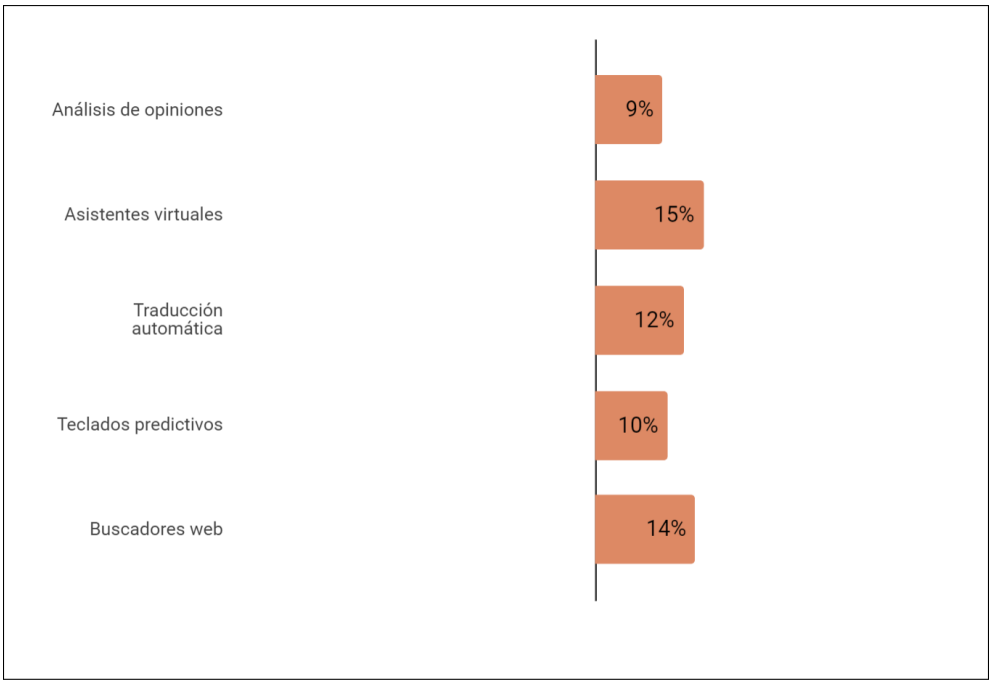

Figura 40: Brecha en la satisfacción de usuario por área de aplicación.

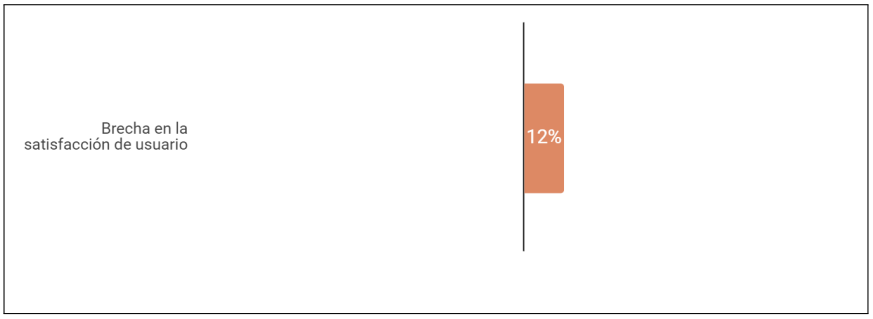

Figura 41: Brecha en la satisfacción de usuario.

Estados Unidos. En general no presenta diferencias estadísticamente significativas entre países, excepto "En precio", donde se muestra un porcentaje significativamente mayor en Estados Unidos que en España. Siri presenta significativamente más limitaciones en Estados Unidos "En desempeño o rendimiento", "Proporciona contenidos ofensivos, tóxicos o inadecuados" y "Proporciona información convincente a primera vista, pero errónea al verificarla". Alexa presenta niveles similares de limitaciones en Estados Unidos y España, excepto en "Sólo ofrece información genérica que ya se conocía" que es significativamente más mencionada en España que en USA. En general ChatGPT presenta niveles similares de limitaciones en Estados Unidos y España. Sin embargo, destaca significativamente en Estados Unidos la percepción de limitaciones "En el desempeño o rendimiento", mientras que en España es significativamente mayor la mención a que "Proporciona información convincente a primera vista, pero errónea al verificarla" y que "No justifica las respuestas ni da acceso a las fuentes que la justifican". Todos los datos se encuentran en en la Tabla 37 del Apéndice D.

En relación a las herramientas de **traducción automática**, se observa una tendencia a reportar menos limitaciones en las herramientas con mayor cantidad de usuarios, tanto en España como en Estados Unidos. Sin embargo, no se observa una tendencia en el tipo de limitaciones reportadas. Google translate presenta diferencias significativas en algunas limitaciones entre España y USA. Destaca con mayor

porcentaje de personas que percibe en Estados Unidos limitaciones "En el desempeño o rendimiento", "En la compatibilidad con otros sistemas o con el equipo o dispositivo", "De seguridad" y "De precio". Mientras que en España es significativamente mayor la percepción de que "Presenta sesgos sistemáticos de traducción y/o estereotipados" y "Traducciones literales que no captura el trasfondo del texto". ChatGPT se percibe en Estados Unidos con significativamente más limitaciones respecto a España: "En el desempeño o rendimiento", "En las funcionalidades" y "De seguridad". Google Bard sólo presenta una limitación con diferencias significativas: en Estados Unidos destaca la limitación frente a España "En la compatibilidad con otros sistemas o con el equipo o dispositivo". Todos los datos se encuentran en en la Tabla 38 del Apéndice D.

En **teclados predictivos**, en España se observan en general menos limitaciones que en Esatos Unidos. Sin embargo, no se observa una tendencia en el tipo de limitaciones. iPhone Keyboard destaca de las demás herramientas porque no presenta diferencias significativas entre las limitaciones percibidas en un país y en el otro. Las funciones predictivas en la redacción de Microsoft Outlook sólo presenta diferencias significativas en la limitación "Compatibilidad con otros sistemas o con el equipo o dispositivo" con más frecuencia en Estados Unidos que en España. En Microsoft 365, en Estados Unidos se percibe significativamente más limitaciones respecto a las observada en España. Destacan "En el desempeño o rendimiento", "En las funcionalidades", "En la compatibilidad con otros sistemas o con el equipo o dispositivo" y "De seguridad". Todos los datos se encuentran en en la Tabla 39 del Apéndice D.

En los **buscadores web** no se observa una tendencia en el tipo de limitaciones que se observan. Google Search destaca como el buscador donde menos usuarios ven limitaciones, tanto en Estados Unidos como en España. En Estados Unidos es significativamente mayor el porcentaje de personas que ve limitaciones "En el desempeño o rendimiento", "En las funcionalidades", "En la compatibilidad con otros sistemas o con el equipo o dispositivo", "De precio", que "Devuelve páginas con contenidos tóxicos, agresivos o inapropiados en general". En España se perciben significativamente mayores las limitaciones "De privacidad" y los "Sesgos / preferencias sistemáticas hacia determinados tipos de páginas" y que "Devuelve páginas relacionadas con la consulta pero que contienen información engañosa". Tanto Bing como Yahoo Search presentan en España significativamente mayores las limitaciones "Sesgos / preferencias sistemáticas hacia determinados tipos de páginas" y que "Devuelve páginas con información previsible con poco valor añadido". También se observa en Bing, una asociación significativamente mayor que en Estados Unidos con la limitación "Devuelve páginas relacionadas con la consulta pero que contienen información engañosa". Todos los datos se encuentran en en la Tabla 40 del Apéndice D.

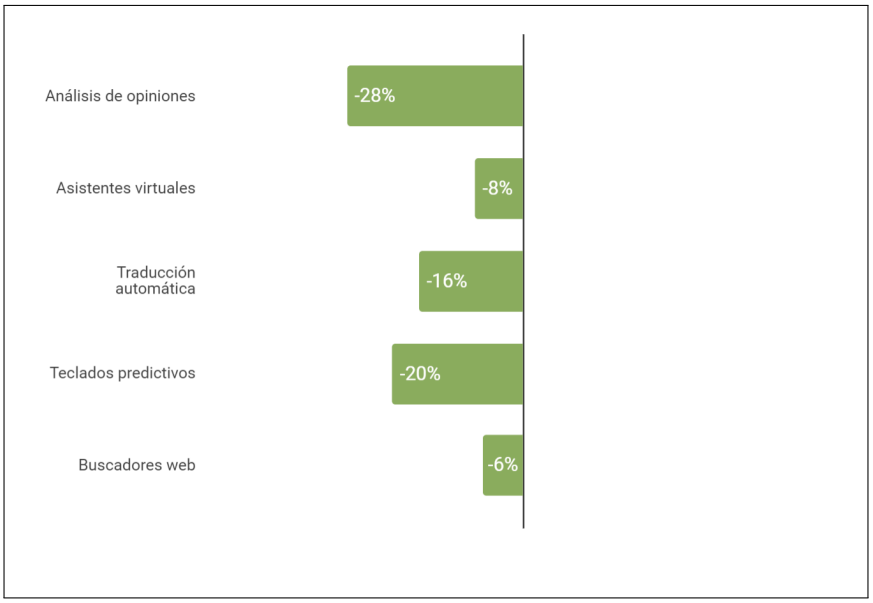

Figura 42: Brecha en las limitaciones de uso por área de aplicación.

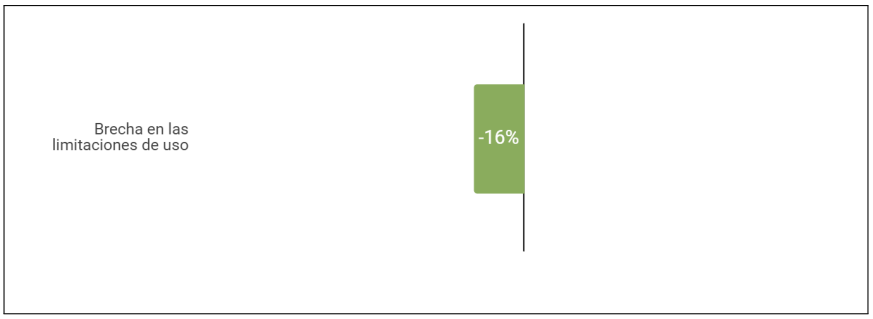

Figura 43: Brecha en las limitaciones de uso.

En la Figura 42 se muestra la brecha en limitaciones de uso de cada una de las áreas de aplicación. Existe una gran diferencia en las limitaciones de uso identificadas entre los usuarios de habla hispana y los de habla inglesa. A pesar de que los de habla inglesa tienen un mayor nivel de satisfacción que los de habla hispana, son quienes indican haber encontrado mayores limitaciones en el uso de las herramientas. Finalmente, como indica la Figura 43 la brecha general en las limitaciones de uso es del 16 % a favor del español, un 9 % menos diferencia que en el Año 1.

Análisis de nuevas dimensiones de evaluación en experiencia de usuario Con motivo de los recientes avances en lo que respecta a modelos de lenguaje generativos, en esta iteración del proyecto se ha llevado a cabo un análisis en profundidad sobre las nuevas dimensiones que surgen en el proceso de evaluación más allá de la efectividad, como son la presencia de sesgos, contenidos dañinos, explicabilidad de resultados, competencias internas del sistema o la informatividad de los resultados. En base a este análisis, se han extendido las preguntas de limitaciones en las encuestas en el ámbito de la experiencia de usuario cubriendo estos aspectos y enunciadas de forma específica para cada área de producto. A continuación, en la figura 44 se analiza en qué dimensiones se observa una mayor brecha lingüística, cuáles son las dimensiones más sensibles para los usuarios en cada área de producto y entre lenguas.

Algunas conclusiones que podrían sugerir estos resultados son las siguientes:

1. Todas las limitaciones se encuentran en un rango de porcentaje de respuestas similar, entre el 10 % y 20 %.
2. Independientemente de la aplicación, los encuestados en inglés identifican considerablemente más limitaciones en términos de efectividad que los encuestados en español.
3. Las limitaciones de sesgo se encuentran en mayor medida en traductores, buscadores y sistemas de opinión, que en teclados predictivos y asistentes.
4. Las respuestas tóxicas o dañinas en buscadores y asistentes (áreas de producto donde se puede aplicar dicha dimensión) no parece ser una limitación especialmente importante para los usuarios en ninguno de los idiomas.
5. La previsibilidad o poca informatividad de los resultados son una limitación para más del 15 % de los encuestados tanto en traductores como en teclados predictivos y buscadores.
6. No hay un patrón común de limitaciones entre idiomas en el caso de las herramientas de opinión. En el resto de aplicaciones parece haber algo de correspondencia, excepto en el caso de la efectividad.

8.4. Conclusiones y evolución de la brecha

En primer lugar, sí que existe, a diferencia del Año 1, una brecha significativa a favor del español en la polaridad reputacional de las opiniones en redes sociales y reseñas en cuatro de las cinco áreas de soluciones, dando lugar a una brecha del -9 % a favor del español.

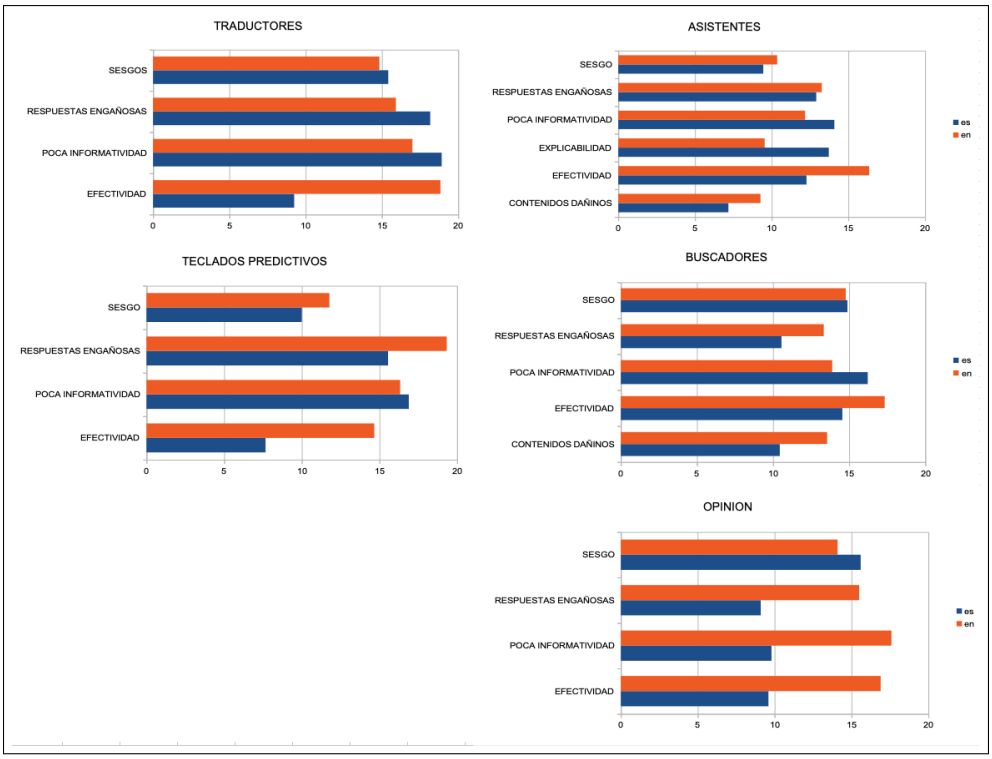

Figura 44: XXX.

En segundo lugar, los aspectos en los que podría mejorarse la propuesta de valor de las soluciones en español siguen siendo el precio y la seguridad/privacidad, mientras que el rendimiento dejó de ser uno de los aspectos a mejorar.

En tercer lugar, aunque en términos generales los hispanos valoran mejor los atributos que los anglosajones (con una brecha del -9 %) y observan menos limitaciones (con una brecha del -16 %), la satisfacción del usuario sigue siendo a favor del inglés (con una brecha del 12 %). Esto hace que exista una brecha global del -6 % en la experiencia de usuario a favor del español, 1 % menos que el Año 1.

9. Agregación de resultados

La Figura 45 resume los valores obtenidos para todos los indicadores obtenidos durante el primer año de proyecto, así como los valores agregados para cada uno de los ámbitos del proyecto. Los valores entre paréntesis representan los indicadores obtenidos en el año 1. Los resultados aparecen subrayados y en negrita en los casos en los que ha habido un aumento de la brecha lingüística. En las siguientes secciones describimos el proceso llevado a cabo para la agregación de indicadores en los distintos ámbitos.

Aunque, como veremos en las siguiente secciones, se dan variaciones a nivel de indicadores específicos, a nivel de ámbitos los resultados son bastante consistentes respecto del año 1. En el ámbito 1 hay una diferencia de 2 puntos, y en los tres ámbitos restantes hay una diferencia de un punto.

9.1. Ámbito 1: Brecha en diseminación y recursos

En cuanto a diseminación y recursos, se mantiene la tendencia observada en la primera iteración del proyecto, con algunas diferencias. Según los resultados de esta iteración el factor más desfavorable es la diseminación, con una brecha en publicaciones (D.1) y proyectos subvencionados (D.2) del 98 % y 96 % respectivamente. En concreto, la brecha en proyectos subvencionados ha ascendido del 88 % al 96 %. Sin embargo, esta volatilidad puede deberse a las pocas muestras de proyectos en español respecto al inglés.

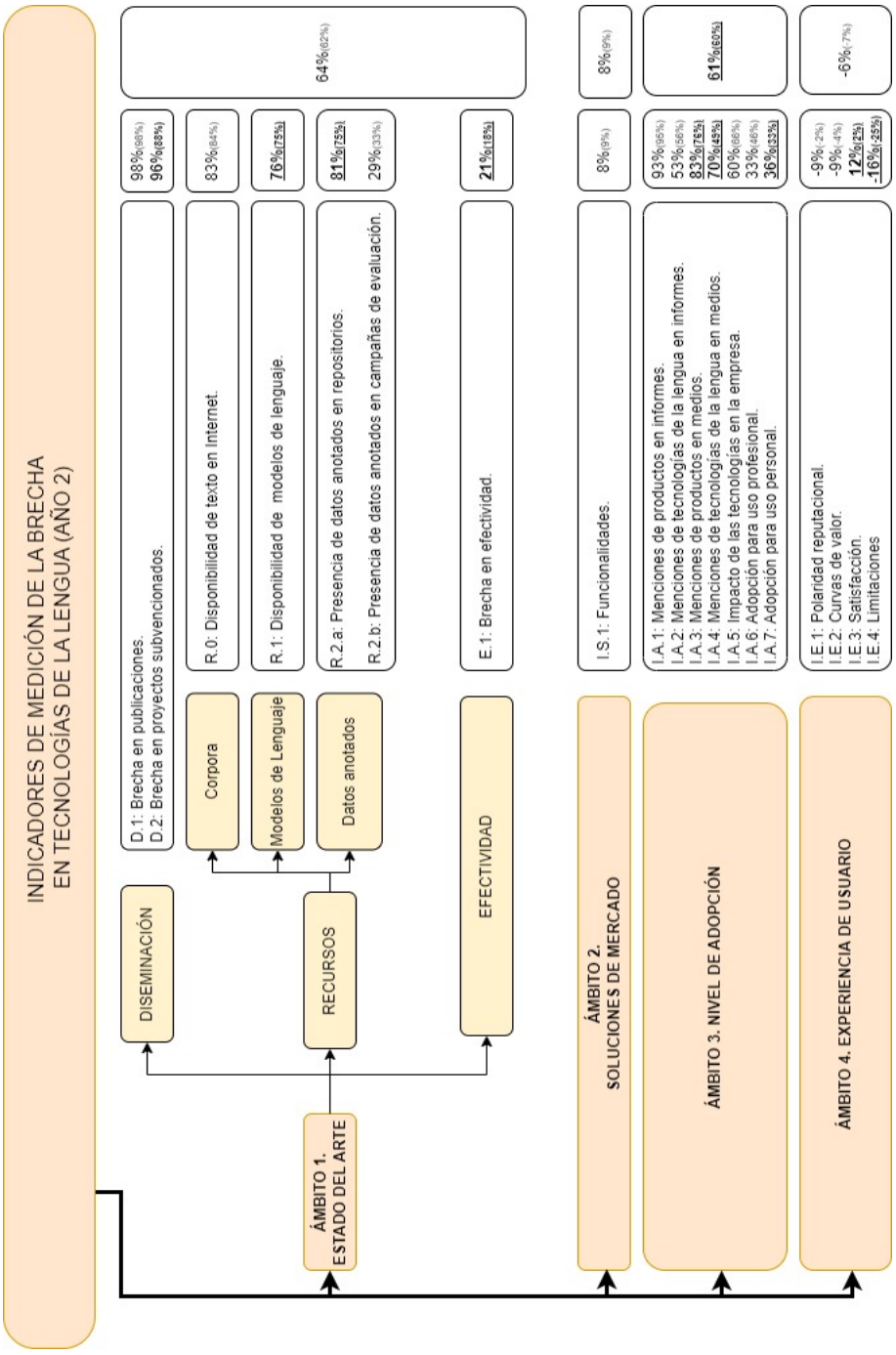

Figura 45: Estimación ODESIA de la brecha entre español e inglés, año 2

En cuanto a recursos, la disponibilidad de textos en internet (R.0) se mantiene estable, como era de esperar dado que no es un indicador susceptible de cambios bruscos. La brecha en disponibilidad de modelos de lenguaje se mantiene muy similar (R.1). Se observa un aumento importante de la disponibilidad de datos anotados en repositorios (R.2), sobre todo debido al incremento de datos para el inglés en Hugging Face. La presencia de datos en campañas de evaluación permanece bastante constante, a pesar de esa diferencia en 4 puntos respecto al año anterior, si tenemos en cuenta el reducido número de muestras y el consiguiente efecto en la volatilidad del indicador R.2.b.

Al igual que en la iteración anterior, la agregación de indicadores se ha planteado como una estimación del esfuerzo necesario para construir una aplicación de Procesamiento del Lenguaje Natural (el campo de la Inteligencia Artificial que nos ocupa) en español, respecto del inglés. Para ello hemos realizado un promedio de los siguientes indicadores de brecha: D.2 (proyectos subvencionados), R.0 (corpora disponible en Internet para desarrollar modelos de lenguaje pre-entrenados), R.1 (modelos de lenguaje disponibles), R.2 (datos anotados disponibles), y E.1 (efectividad de los modelos de lenguaje en tareas de PLN usando datos comparables para el fine-tuning). Nótese que para hacer este promedio hemos descartado el indicador de brecha de publicaciones científicas, ya que no podemos establecer un impacto directo en el coste de desarrollar aplicaciones eficaces y eficientes, dado que la algorítmica subyacente en el estado del arte es independiente de la lengua.

9.2. Ámbito 1: Brecha en efectividad

Para la estimación del indicador de efectividad se han empleado 15 datasets, de los cuales 10 pertenecen al leader board desarrollado en el proyecto con datos de test no públicos, y 5 de ellos suponen una extesión del estudio de la brecha con datos públicos. La tabla 24 muestra las brechas obtenidas en cada uno de los datasets sobre los que se ha realizado el estudio.

En cuanto a tareas abstractas se refiere, para la estimación de la brecha, se ha cubierto la clasificación binaria (EXIST 2022 tarea 1, EXIST 2023 tarea 1, DIPROMATS 2023 tarea 1), clasificación multiclase, jerárquica y/o multilabel (EXIST 2022 tarea 2, EXIST 2023 tareas 2 y 3, DIPROMATS 2023 tareas 2 y 3), *learning with disagreement* (EXIST 2023, tareas 1,2 y 3), regresión (STS 2017), etiquetado de secuencias (DIANN). Dentro de lo que se consideran problemas dinámicos, en la estimación de la brecha consideramos el *question answering* con anotación de secuencias (SQUAD/SQAC 2024).

Los dominios cubiertos en el segundo año en el cálculo de la brecha en efectividad incluyen: geopolítica (DIPROMATS), biomedicina (DIANN), publicaciones académicas (SQUAD/SQAC), y análisis de redes sociales, en concreto, identificación de sexismo (EXISTS). Además, para la experimentación se han incluido tres datasets adicionales de dominio público que incluyen textos de dominio periodístico (MLDOC), conocimiento enciclopédico y consultas en buscadores (MULTICONER) y resolución de similitud textual (STS).

Es interesante observar que, a pesar de la diversidad de datasets y tareas, en todos los casos menos uno (EXIST 2023 tarea 2, donde el rendimiento en ambos idiomas es similar) se ha medido una brecha favorable al inglés, lo que proporciona una fuerte evidencia estadística de la existencia de esa brecha. Puede apreciarse, también, que hay un valor muy diferente al resto en el caso de la tarea DIPROMATS Task 3. Esa tarea es la más difícil de las contempladas en nuestra experimentación, ya que es un problema de clasificación multiclase y multilabel con 13 clases distribuidas de forma muy desigual. Al ser la más difícil, también es la tarea en la que los modelos de lenguaje aportan una mejora más sustancial respecto a las aproximaciones baseline sin conocimiento lingüístico. Aunque es un resultado en el que merece la pena profundizar, a efectos del cálculo del gap lo hemos descartado por razones estadísticas, ya que se trata de un outlier ($p < 0,01$ según el test de Grubbs).

Una vez eliminado el outlier, el indicador de brecha de efectividad EF,1 se ha calculado según se describe en la sección 4.1.6 sobre las otras catorce tareas. **El resultado final (con su error estándar) es una estimación de la brecha porcentual del 20 $\pm$ 06 a favor del inglés.** Aunque hay una variación apreciable entre tareas, el error estándar nos indica que, en cualquier caso, la brecha real promedio estará en una horquilla entre el 14 % y el 26 %. Estos datos son compatibles con los obtenidos el primer año.

Esta brecha está medida exclusivamente sobre modelos discriminativos. Nuestra experimentación

Tabla 24: Descripción de tareas y modelos

TAREA	Mejor LLM ES	Tipo de Tarea	Brecha
CORE TASKS			
EXIST 2022 Task 1	dccuchile/bert-base-spanish-wwm-cased	clasificación binaria	17 %
EXIST 2022 Task 2	PlanTL-GOB-ES/roberta-large-bne	clasificación multiclase	10 %
DIPROMATS 2023 Task 1	dccuchile/bert-base-spanish-wwm-cased	Clasificación binaria	11 %
DIPROMATS 2023 Task 2	xlm-roberta-large	clasificación multiclase multilabel	48 %
DIPROMATS 2023 Task 3	xlm-roberta-large	Clasificación multiclase multilabel	293 %
DIANN 2023 Task 1	PlanTL-GOB-ES/roberta-base-bne	Etiquetado de secuencias	0.4 %
SQUAD/SQAC 2024	PlanTL-GOB-ES/roberta-large-bne	Etiquetado	19 %
EXIST 2023 Task 1	xlm-roberta-large	clasificación binaria learning with disagreement	10 %
EXIST 2023 Task 2	xlm-roberta-large	clasificación jerárquica multiclase learning with disagreement	-3 %
EXIST 2023 Task 3	xlm-roberta-large	clasificación multiclase multilabel jerarquica learning with disagreement	12 %
EXTENDED TASKS			
DIANN Task 2	dccuchile/bert-base-spanish-wwm-cased	etiquetado de secuencias	72 %
MLDoc	PlanTL-GOB-ES/roberta-base-bne	clasificación multiclase	40 %
MultiCoNER CoNLL 2022	xlm-roberta-large	etiquetado de secuencias	5 %
STS-2017	dccuchile/bert-base-spanish-wwm-cased	regresión	15 %
SQUAD/SQAC 2016	xlm-roberta-large	etiquetado de secuencias	25 %

con modelos generativos ha dado lugar a primeras estimaciones de brecha que no hemos incluido en el indicador global. Nuestra experimentación inicial con modelos generativos no tiene suficiente representatividad como para incluirla en el cálculo de la brecha, pero indica, provisionalmente, que **en los modelos generativos no hay una brecha mayor que la que hemos medido con los discriminativos**:

- GPT-4 (el modelo más potente junto con Claude 3 en el momento de finalizar nuestra experimentación) en modo zero-shot, aplicado a tres tareas discriminativas de nuestro leaderboard, arroja una brecha promedio del 18 % entre el español y el inglés, muy similar a nuestro resultado en modelos discriminativos.
- En nuestro dataset UNED ACCESO de preguntas de exámenes de acceso a la universidad, los modelos generativos abiertos LLama-2, Gemma y Mistral tienen una brecha promedio del 12 % entre el español y el inglés. Y los modelos más potentes (Claude-3 y GPT-4) dan una brecha ligeramente negativa (-2 % en ambos casos), es decir, se comportan un poco mejor en español que en inglés. Teniendo en cuenta que las preguntas del dataset están originalmente en español (y son traducidas manualmente al inglés), es muy posible que los modelos hayan visto parte de las respuestas en su fase de entrenamiento (contaminación), y tampoco es descartable que haya algún artefacto de traducción (aunque son traducciones manuales realizadas por profesionales). Descartando estos efectos, no sería raro que la medición del gap estuviera en niveles parecidos a los del experimento anterior con GPT-4. En cualquier caso, necesitamos profundizar en nuestro diseño experimental para obtener cifras fiables.

9.3. Ámbitos de implantación

En cuanto a los ámbitos de implantación, es decir, los ámbitos de soluciones de mercado, nivel de adopción y experiencia de usuario, los resultados en promedio para cada uno de ellos son muy similares a los obtenidos en el año anterior, distanciándose la brecha en un punto en todos los casos.

La brecha en soluciones de mercado se mantiene en un 8 % descendiendo solo un punto respecto al año anterior. En cuanto a la brehca en nivel de adopción, aparecen variaciones a nivel de indicador específico, aunque el promedio se mantiene prácticamente constante. En concreto, ascienden las menciones de productos en medios (I.A.3) de un 76 % a un 83 %, las menciones de tecnologías de la lengua en medios

de un 49 % a un 70 % (I.A.4) y la brecha en adopción para uso personal (I.A.7) que asciende de un 33 % a un 36 %. Sin embargo, descienden la brecha en menciones de productos informes (I.A.1) de un 95 % a un 93 %, las menciones de tecnologías en informes (I.A.2) de un 56 % a un 53 %, el impacto de las tecnologías en la empresa (I.A.5) de un 66 % a un 60 % y la adopción para uso profesional (I.A.7) de un 46 % a un 33 %.

En cuanto al ámbito de experiencia de usuario, también se mantiene constante en promedio, aunque hay variaciones a nivel de indicador específico. Se ha obtenido el mismo patrón que el año anterior. Los indicadores de polaridad reputacional (I.E.1) y curvas de valor (I.E.2) donde los indicadores se estiman a partir de opiniones en la web, aparece una brecha negativa en favor del español. Lo mismo ocurre en el indicador de limitaciones (I.E.4) en donde se encuesta a individuos sobre las deficiencias específicas de los productos analizados. Sin embargo, en el caso de las encuestas de satisfacción (I.E.3) los usuarios de tecnologías en inglés se sienten más satisfechos, efecto que ha crecido en los resultados de este año (12 % frente al 2 % obtenido en el año anterior). Esto se compensa con la reducción de brecha en favor del español en los otros tres indicadores (-9 % frente a 2 %, -9 % frente a -4 % y 16 % frente a 25 %).

De manera adicional, se han ampliado las encuestas sobre limitaciones para identificar aspectos de la calidad de las tecnologías definidos en la sección 3 de este documento.

9.4. Conclusiones

En este segundo año de proyecto hemos medido, de nuevo, una brecha significativa en casi todos los ámbitos estudiados, lo que confirma la necesidad de impulsar la IA en español como parte de cualquier estrategia nacional de desarrollo tecnológico.

Entre el primer año y el segundo se ha producido la irrupción definitiva de la IA generativa, y esperábamos ver cambios en los indicadores analizados. Sin embargo, aunque la investigación, el desarrollo y la implantación se han acelerado enormemente desde la aparición de ChatGPT el 30 de noviembre de 2022, parecen haberlo hecho de forma paralela en el mundo angloparlante y en el hispanoparlante, de tal manera que la brecha se ha mantenido, promediando sobre todos los ámbitos de análisis, similar a la que había antes de ChatGPT.

Es destacable el esfuerzo realizado para la medición de la brecha en efectividad de los modelos de lenguaje: se han desarrollado en el ámbito del proyecto nueve datasets, se ha experimentando con 14 modelos de lenguaje discriminativos abiertos, tres modelos generativos propietarios (Claude 3, GPT-4 y GPT-3.5) y tres modelos generativos abiertos (Mistral, Llama-2 y Gemma). En la experimentación se han realizado más de 3,500 procesos de fine-tuning y evaluación sobre cada una de las tareas discriminativas en cada uno de los idiomas. Este esfuerzo sienta las bases para un seguimiento continuado de los modelos de lenguaje aplicables al español, además de servir para estimar la brecha inglés-español.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE TRANSFORMACIÓN DIGITAL
red.es
Plan de Recuperación, Transformación y Resiliencia
UNED

Agradecimientos

Este trabajo ha sido financiado por la Unión Europea - NextGenerationEU a través del "Plan de Recuperación, Transformación y Resiliencia", por el Ministerio de Asuntos Económicos y Transformación Digital y por la UNED. Sin embargo, los puntos de vista y las opiniones expresadas son únicamente los del autor o autores y no reflejan necesariamente los de la Unión Europea o la Comisión Europea. Ni la Unión Europea ni la Comisión Europea pueden ser consideradas responsables de los mismos.

Bibliografía

Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2017. QUINT: Interpretable question answering over knowledge bases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 61–66, Copenhagen, Denmark. Association for Computational Linguistics.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects. ArXiv:2309.07445 [cs].

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. pages 5–14.

Enrique Amigo and Agustín Delgado. 2022. Evaluating extreme hierarchical multi-label classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5819, Dublin, Ireland. Association for Computational Linguistics.

Enrique Amigó, Stefano Mizzaro, and Damiano Spina. 2022. Ranking interruptus: When truncated rankings are better and how to measure that. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 588–598, New York, NY, USA. Association for Computing Machinery.

Enrique Amigó, Damiano Spina, and Jorge Carrillo de Albornoz. 2018. An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 625–634. ACM.

Enrique Amigó and Agustín Delgado. 2022. Evaluating extreme hierarchical multi-label classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume Volume 1: Long Papers, page 5809–5819, Dublin, Ireland. ACL.

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023. How do in-context examples affect compositional generalization? *CoRR*, abs/2305.04835.

Anjum and Rahul Katarya. 2023. Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. *Int. J. Inf. Secur.*, 23(1):577–608.

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John T. Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *CoRR*, abs/1909.03012.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021a. Program synthesis with large language models.

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. 2021b. Program synthesis with large language models. *CoRR*, abs/2108.07732.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Marco Baroni. 2019. Linguistic generalization and compositionality in modern artificial neural networks. *CoRR*, abs/1904.00157.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Jasmijn Bastings, Marco Baroni, Jason Weston, Kyunghyun Cho, and Douwe Kiela. 2018. Jump to better conclusions: SCAN both left and right. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 47–55, Brussels, Belgium. Association for Computational Linguistics.

Gregor Betz, Christian Voigt, and Kyle Richardson. 2021. Critical thinking for language models. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 63–75, Groningen, The Netherlands (online). Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Daniel M. Bikel and Imed Zitouni. 2012. *Multilingual Natural Language Processing Applications: From Theory to Practice*. IBM Press.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Boost.ai. 2023. Gartner® Magic Quadrant™ for Enterprise Conversational AI Platforms. https://www.boost.ai/reports/gartner-magic-quadrant-for-enterprise-conversational-ai-platforms. [Accessed 30-11-2023].

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA@CLiC-it*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3497–3507, Brussels, Belgium. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017a. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017b. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2023. Creativity support in the age of large language models: An empirical study involving emerging writers. *ArXiv*, abs/2309.12570.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of semantic similarity—a survey. *ACM Comput. Surv.*, 54(2).

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 621–630, New York, NY, USA. Association for Computing Machinery.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 13–22, New York, NY, USA. Association for Computing Machinery.

Honghua Chen and Nai Ding. 2023. Probing the creativity of large language models: Can models produce divergent semantic association?

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. *CoRR*, abs/2002.05867.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Counterpoint Technology Market Research. 2023. Global Smartphone Shipments Market Data (Q4 2021 – Q3 2023). https://www.counterpointresearch.com/global-smartphone-share/. [Accessed 27-11-2023].

Danilo Croce, Daniele Rossini, and Roberto Basili. 2019. Auditing deep learning processes through kernel-based explanatory models. 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)*, pages 4037–4046, Hong Kong, China. Association for Computational Linguistics.

David Curry. 2023. Android Statistics (2023). https://www.businessofapps.com/data/android-statistics. [Accessed 27-02-2023].

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Ernest Davis. 1990. *Representations of Commonsense Knowledge*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Bhuwan Dhingra, Manaal Faruqui, Ankur P. Parikh, Ming-Wei Chang, Dipanjan Das, and William W. Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. *CoRR*, abs/1906.01081.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Percy Liang et al. 2023. Holistic evaluation of language models.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Hermenegildo Fabregat, Juan Martínez-Romo, and Lourdes Araujo. 2018. Overview of the DIANN task: Disability annotation task. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 1–14. CEUR-WS.org.

Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms).

Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. 2023. Explaining how transformers use context to build predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5513, Toronto, Canada. Association for Computational Linguistics.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, Online. Association for Computational Linguistics.

Jerry A. Fodor and Ernest Lepore, editors. 2002. *Compositionality Papers*. Oxford University Press UK.

Giorgio Franceschelli and Mirco Musolesi. 2023. On the creativity of large language models.

E. Frank, M. A. Hall, G. Holmes, R. Kirkby, B. Pfahringer, and I. H. Witten. 2005. *Weka: A machine learning workbench for data mining.*, pages 1305–1314. Springer, Berlin.

Gartner, Inc. 2023. Magic Quadrant for Insight Engines. https://www.gartner.com/doc/reprints?id=1-2C0HFZ80&ct=221215&st=sb. [Accessed 30-11-2023].

Mouzhi Ge, Carla Delgado, and Dietmar Jannach. 2010. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. pages 257–260.

J. Ghosh. 2003. Scalable clustering methods for data mining. In Nong Ye, editor, *Handbook of Data Mining*. Lawrence Erlbaum.

Statista GmbH. 2023. Market share of leading desktop search engines worldwide from January 2015 to July 2023. https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/. [Accessed 30-11-2023].

Fréderic Godin, Kris Demuynck, Joni Dambre, Wesley De Neve, and Thomas Demeester. 2018. Explaining character-aware neural networks for word-level prediction: Do they discover linguistic rules? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3275–3284. Association for Computational Linguistics.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 166–175, New York, NY, USA. Association for Computing Machinery.

Emily Goodwin, Koustuv Sinha, and Timothy J. O'Donnell. 2020. Probing linguistic systematicity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1958–1969, Online. Association for Computational Linguistics.

greatcontent GmbH. 2023. The 11 Best Machine (AI) Translation Tools to Use in 2023. https://greatcontent.com/machine-ai-translation-tools/. [Accessed 30-11-2023].

Yu Gu, Sue Kase, Michelle Vanni, Brian M. Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2020. Beyond I.I.D.: three levels of generalization for question answering on knowledge bases. *CoRR*, abs/2011.07743.

Yuling Gu, Bhavana Dalvi Mishra, and Peter Clark. 2023. Do language models have coherent mental models of everyday things? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1892–1913, Toronto, Canada. Association for Computational Linguistics.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2021. Maria: Spanish language models. *arXiv preprint arXiv:2107.07253*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. Folio: Natural language reasoning with first-order logic.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097, Toronto, Canada. Association for Computational Linguistics.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world state with recurrent entity networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. ArXiv:2009.03300 [cs].

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohammad Hossin and Sulaiman M.N. 2015. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5:01–11.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Yi-Chong Huang, Xia-Chong Feng, Xiao-Cheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *CoRR*, abs/2104.14839.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Fanchao Qi, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *ArXiv*, abs/2305.08322.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2021. Compositionality decomposed: How do neural networks generalise? (extended abstract). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Insider Inc. 2023a. IBEX 35 Market capitalization. https://markets.businessinsider.com/index/market-capitalization/ibex_35. [Accessed 27-10-2023].

Insider Inc. 2023b. S&P 500 Market capitalization. https://markets.businessinsider.com/index/market-capitalization/s&p_500. [Accessed 07-02-2024].

Matthew Izatt. 2022. Year in Review: 12 awesome ways for developers to learn, build, and grow with Google Workspace. https://developers.googleblog.com/2022/12/year-in-review-12-awesome-ways-for-developers-to-learn-build-grow-google-workspace.html. [Accessed 30-11-2023].

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Justin Johnson and Taghi Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6:27.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2023. *Speech and Language Processing. Third Edition draft*.

Aishwarya Kamath and Rajarshi Das. 2018. A survey on semantic parsing. *ArXiv*, abs/1812.00978.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring compositional generalization: A comprehensive method on realistic data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8082–8090. AAAI Press.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9087–9105. Association for Computational Linguistics.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020a. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020b. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy Liang. 2019. Spoc: Search-based pseudocode to code. *CoRR*, abs/1906.04908.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Brenden M. Lake and Marco Baroni. 2017. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2016. Building machines that learn and think like people. *CoRR*, abs/1604.00289.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Minhyeok Lee. 2023. A mathematical investigation of hallucination and creativity in gpt models. *Mathematics*, 11(10).

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2021. A comprehensive comparative evaluation and analysis of distributional semantic models. *CoRR*, abs/2105.09825.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021a. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.

Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. 2021b. Guided generation of cause and effect. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Chao-Chun Liang, Shih-Hong Tsai, Ting-Yun Chang, Yi-Chung Lin, and Keh-Yih Su. 2016. A meaning-based English math word problem solver with understanding, reasoning and explanation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 151–155, Osaka, Japan. The COLING 2016 Organizing Committee.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2023. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Cinny Little. 2021. The Forrester New Wave™: AI-Enabled Consumer Intelligence Platforms, Q3 2021. https://www.forrester.com/report/The-Forrester-New-Wave-AIEnabled-Consumer-Intelligence-Platforms-Q3-2021/RES161546. [Accessed 08-11-2023].

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. Semeval-2022 task 11: Multilingual complex named entity recognition (multiconer). In *Proceedings of the 16th international workshop on semantic evaluation (SemEval-2022)*, pages 1412–1437.

Guillermo Marco, Julio Gonzalo, and Luz Rello. 2022. A systematic evaluation of the creative writing skills of transformer deep neural networks. Technical report. 10.2139/ssrn.4042578; 10.2139/ssrn.4042578.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Marina Meilă. 2007. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27.

Pablo Moral, Guillermo Marco Remón, Julio Gonzalo Arroyo, Jorge Carrillo-de Albornoz, and Iván Gonzalo-Verdugo. 2023-09. Overview of dipromats 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Tomoko Murakami, Koichiro Mori, and Ryohei Orihara. 2008. Metrics for evaluating the serendipity of recommendation lists. In *New Frontiers in Artificial Intelligence*, pages 40–46, Berlin, Heidelberg. Springer Berlin Heidelberg.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

Santiago Ontañón, Joshua Ainslie, Vaclav Cvicek, and Zachary Kenneth Fisher. 2022. Logicinference: A new dataset for teaching logical inference to seq2seq models. *ArXiv*, abs/2203.15099.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

PeerSpot. 2023. Best Chatbot Development Platforms. `https://www.peerspot.com/categories/chatbot-development-platforms`. [Accessed 27-11-2023].

Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating robustness and interpretability of link prediction via adversarial modifications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3336–3347, Minneapolis, Minnesota. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Yejin Choi, and Zaïd Harchaoui. 2021. MAUVE: human-machine divergence curves for evaluating open-ended text generation. *CoRR*, abs/2102.01454.

Laura Plaza, Jorge Carrillo de Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2023. Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization. Experimental IR Meets Multilinguality, Multimodality, and Interaction. In *Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

Jing Qian, Hong Wang, Zekun Li, Shiyang Li, and Xifeng Yan. 2023. Limitations of language models in arithmetic and symbolic induction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9285–9298, Toronto, Canada. Association for Computational Linguistics.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal commonsense reasoning in dialog. 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)*, pages 7066–7076, Online. Association for Computational Linguistics.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. WebCPM: Interactive web search for Chinese long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8968–8988, Toronto, Canada. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia

Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016a. Squad: 100,000+ questions for machine comprehension of text.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016b. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Tony Redmond. 2022. Office 365 Number of Users Reaches 345 Million Paid Seats. https://office365itpros.com/2022/04/28/office-365-number-of-users/. [Accessed 30-11-2023].

Shuo Ren, Daya Guo, Shuai Lu, Long Zhou, Shujie Liu, Duyu Tang, Neel Sundaresan, Ming Zhou, Ambrosio Blanco, and Shuai Ma. 2020. Codebleu: a method for automatic evaluation of code synthesis. *CoRR*, abs/2009.10297.

Mohammed Saeed, Nicola De Cao, and Paolo Papotti. 2023. Querying large language models with sql.

Swarnadeep Saha, Yixin Nie, and Mohit Bansal. 2020. ConjNLI: Natural language inference over conjunctive sentences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8240–8252, Online. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Samsung. 2022. Samsung Further Develops Bixby, Introducing a New Language and Setting a Foundation for Future Growth. https://news.samsung.com/global/samsung-further-develops-bixby-introducing-a-new-language-and-setting-a-foundation-fo [Accessed 27-11-2023].

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. *ArXiv*, abs/1904.01557.

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? 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)*, pages 922–938, Online. Association for Computational Linguistics.

Rohit Shewale. 2023a. 30+ Google Bard Statistics 2023 (Trends & Demographics). https://www.demandsage.com/google-bard-statistics/. [Accessed 27-11-2023].

Rohit Shewale. 2023b. 58 Gmail Statistics For 2023 (Worldwide Demographics). https://www.demandsage.com/gmail-statistics/. [Accessed 30-11-2023].

Rohit Shewale. 2023c. ChatGPT Statistics — Detailed Insights On Users (2023). https://www.demandsage.com/chatgpt-statistics/. [Accessed 27-11-2023].

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2019. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. *CoRR*, abs/1912.01734.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gino Silva-Payne. 2022. Gmail Vs Outlook: Which Is The Best In 2023? — emailmeter.com. https://www.emailmeter.com/blog/gmail-vs-outlook. [Accessed 30-11-2023].

Similarweb LTD. 2023. Website Traffic - Check and Analyze Any Website. https://www.similarweb.com/. [Accessed 30-11-2023].

Craig Smith. 2023. Amazon Alexa Statistics and User Count (2023). https://expandedramblings.com/index.php/amazon-alexa-statistics/. [Accessed 27-11-2023].

Jeff Speaks. 2021. Theories of Meaning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2021 edition. Metaphysics Research Lab, Stanford University.

Aarohi et al Srivastava. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. ArXiv:2206.04615 [cs, stat].

Robert Stalnaker. 1999. *Context and Content*. Oxford University Press, Oxford, UK.

M. Steinbach, G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *CoRR*, abs/1904.01172.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Douglas Summers-Stay, Clare R. Voss, and Stephanie M. Lukin. 2023. Brainstorm, then select: a generative language model improves its creativity score. In *The AAAI-23 Workshop on Creative AI Across Modalities*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. ArXiv:2210.09261 [cs].

Alona Sydorova, Nina Poerner, and Benjamin Roth. 2019. Interpretable question answering on knowledge bases and text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4943–4951, Florence, Italy. Association for Computational Linguistics.

Zoltán Gendler Szabó. 2022. Compositionality. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2022 edition. Metaphysics Research Lab, Stanford University.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE TRANSFORMACIÓN DIGITAL
red.es
Plan de Recuperación, Transformación y Resiliencia
UNED

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021b. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Josef Valvoda, Naomi Saphra, Jonathan Rawski, Adina Williams, and Ryan Cotterell. 2022. Benchmarking compositionality with formal languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6007–6018, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Noortje J. Venhuizen, Petra Hendriks, Matthew W. Crocker, and Harm Brouwer. 2019. A framework for distributional formal semantics. In *Logic, Language, Information, and Computation*, pages 633–646, Berlin, Heidelberg. Springer Berlin Heidelberg.

Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76.

Nikos Voskarides, Edgar Meij, Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. 2015. Learning to explain entity relationships in knowledge graphs. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 564–574, Beijing, China. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018a. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.

Lihan Wang, Bowen Qin, Binyuan Hui, Bowen Li, Min Yang, Bailin Wang, Binhua Li, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022. Proton: Probing schema linking information from pre-trained language models for text-to-sql parsing. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 1889–1898, New York, NY, USA. Association for Computing Machinery.

Su Wang, Greg Durrett, and Katrin Erk. 2018b. Modeling semantic plausibility by injecting world knowledge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 303–308, New Orleans, Louisiana. Association for Computational Linguistics.

Su Wang, Greg Durrett, and Katrin Erk. 2018c. Modeling semantic plausibility by injecting world knowledge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 303–308, New Orleans, Louisiana. Association for Computational Linguistics.

Wei Wang, Chen Wu, and Ming Yan. 2018d. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1705–1714. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 602–631, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021. Towards quantifiable dialogue coherence evaluation. 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)*, pages 2718–2729, Online. Association for Computational Linguistics.

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206, Berlin, Germany. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Heyang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, Youxuan Jiang, Michihiro Yasunaga, Sungrok Shim, Tao Chen, Alexander Fabbri, Zifan Li, Luyao Chen, Yuwen Zhang, Shreya Dixit, Vincent Zhang, Caiming Xiong, Richard Socher, Walter Lasecki, and Dragomir Radev. 2019. CoSQL: A conversational text-to-SQL challenge towards cross-domain natural language interfaces to databases. 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)*, pages 1962–1979, Hong Kong, China. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

ChengXiang Zhai. 2008. Book review: Statistical language models for information retrieval by chengxiang zhai. *Computational Linguistics*, 36:279–281.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Y. Zhao and G. Karypis. 2001. Criterion functions for document clustering: Experiments and analysis. Technical Report TR 01–40, Department of Computer Science, University of Minnesota, Minneapolis, MN.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. ArXiv:2304.06364 [cs].

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371, Online. Association for Computational Linguistics.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji rong Wen. 2023. Large language models for information retrieval: A survey. *ArXiv*, abs/2308.07107.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools.

Julia El Zini and Mariette Awad. 2022. On the explainability of natural language processing deep models. *ACM Comput. Surv.*, 55(5).

Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 88–109, Harbin, China. Chinese Information Processing Society of China.

ZoomInfo Technologies LLC. 2023. ZoomInfo: Go-to-Market Software | Scale & Power Your GTM. `https://www.zoominfo.com`. [Accessed 08-11-2023].

Keneilwe Zuva and Tranos Zuva. 2017. Diversity and serendipity in recommender systems. In *Proceedings of the International Conference on Big Data and Internet of Thing*, BDIOT2017, page 120–124, New York, NY, USA. Association for Computing Machinery.

A. Bolsa de expresiones de tecnologías de la lengua

Tabla 25: Expresiones regulares de tecnologías de la lengua.

TECNOLOGÍA	INGLÉS	ESPAÑOL	FUENTE
Linguística de corpus	annotated corpora	corpus anotados	LREC Keywords
	annotated corpus	corpus anotado	LREC Keywords
	corpora annotation	anotacion de corpus	LREC Keywords
	corpus annotation		LREC Keywords
	corpus linguistics	linguistica de corpus	LREC Keywords
	english corpora	corpus en ingles	LREC Keywords
	english corpus		LREC Keywords
	spanish corpora	corpus en español	LREC Keywords
	spanish corpus		LREC Keywords
	text corpora	corpus de texto	LREC Keywords
	text corpus		LREC Keywords
	treebank		(Bikel and Zitouni, 2012)
	treebanks		(Bikel and Zitouni, 2012)
Sistemas de diálogo	spoken dialog system	sistema de dialogo hablado	(Bikel and Zitouni, 2012)
	spoken dialog systems	sistemas de dialogo hablado	(Bikel and Zitouni, 2012)
	chatbot		(Jurafsky and Martin, 2023)
	chatbots		(Jurafsky and Martin, 2023)
	dialogue modeling	modelado de dialogo	
		modelado de dialogos	TACL Areas of Interest
Semántica del discurso	argument mining	mineria de argumentos	Wikipedia - Common NLP tasks
	implicit semantic role labelling	etiquetado implicito de roles semanticos	Wikipedia - Common NLP tasks
	topic boundary detection	deteccion de limite de tema	(Bikel and Zitouni, 2012)
Menciones de PLN no específicas	computational linguistics	linguistica computacional	Wikipedia
	language technologies	tecnologias de lenguaje	
		tecnologias del lenguaje	Wikipedia
	language technology	tecnologia de lenguaje	
		tecnologia del lenguaje	Wikipedia
	natural language processing	procesamiento de lenguaje natural	
		procesamiento del lenguaje natural	Wikipedia
Sistemas generativos	natural language generation	generacion de lenguaje natural	
		generación de lenguajes naturales	LREC Keywords
	text to image generation	generacion de imagenes a partir de texto	
		generacion de imagenes a partir de textos	Wikipedia - Common NLP tasks
	text to scene generation	generacion de escenas a partir de texto	
		generacion de escenas a partir de textos	Wikipedia - Common NLP tasks
	text to video generation	generacion de video a partir de texto	
	generacion de videos a partir de texto		
	generacion de video a partir de textos		
		generacion de videos a partir de textos	Wikipedia - Common NLP tasks
Modelado de lenguaje	fill mask model	modelo fill mask	Hugging Face Tasks
	fill mask models	modelos fill mask	Hugging Face Tasks
	sentence similarity model	modelo de similitud de oraciones	
		modelo de similitud de frases	Hugging Face Tasks
	sentence similarity models	modelos de similitud de oraciones	
		modelos de similitud de frases	Hugging Face Tasks
	language based ai	ai basado en lenguaje	TACL Areas of Interest
	language modeling	modelado de lenguaje	
		modelado del lenguaje	TACL Areas of Interest
	word embedding		TACL Areas of Interest
	word embeddings		TACL Areas of Interest
	language model	modelo de lenguaje	LREC Keywords
	language models	modelos de lenguaje	LREC Keywords
Semántica léxica	terminology extraction	extractores de terminologia	Wikipedia - Common NLP tasks
	word sense disambiguation	desambiguacion linguistica	Wikipedia - Common NLP tasks
	morphological induction		B(Bikel and Zitouni, 2012)
Análisis morfológico	morphology induction	inducción morfologica	(Bikel and Zitouni, 2012)
	lemmatization	lematizacion	Wikipedia - Common NLP tasks
	lemmatisation		Wikipedia - Common NLP tasks
	morphological segmentation	segmentacion morfologica	Wikipedia - Common NLP tasks
	part of speech tagging	etiquetado de parte del discurso	
		etiquetado lexico	Wikipedia - Common NLP tasks

	pos tagging		Wikipedia - Common NLP tasks
Reconocimiento de entidades	entity linking	enlazamiento de entidades	Wikipedia - Common NLP tasks
	named entity recognition	reconocimiento de entidades nombradas	Wikipedia - Common NLP tasks
Comprensión del lenguaje natural	deep linguistic processing	procesamiento linguistico profundo	Wikipedia
	natural language programming	programacion de lenguaje natural	
		programacion del lenguaje natural	Wikipedia
	natural language understanding	comprension de lenguaje natural	
		comprension del lenguaje natural	Wikipedia
Reconocimiento óptico de caracteres	optical character recognition	reconocimiento optico de caracteres	Wikipedia - Common NLP tasks
Respuesta a preguntas	question answering model	modelo de respuesta a preguntas	Hugging Face Tasks
	question answering models	modelos de respuesta a preguntas	Hugging Face Tasks
	automatic question answering	respuesta automatica a preguntas	
		respuesta automatica de preguntas	Wikipedia - Common NLP tasks
Semántica relacional	computational semantics	semántica computacional	(Jurafsky and Martin, 2023)
	semantic parsing	parseo semantico	Wikipedia - Common NLP tasks
	semantic role labelling	etiquetado de roles semanticos	Wikipedia - Common NLP tasks
Análisis de sentimientos	opinion mining	mineria de opinion	
		mineria de opiniones	TACL Areas of Interest
	sentiment analysis	analisis de sentimiento	Wikipedia - Common NLP tasks
	sentiment classification	clasificacion de sentimiento	Wikipedia - Common NLP tasks
Lingüística estadística	statistical linguistics	linguistica estadistica	LREC Keywords
	quantitative linguistics	linguistica cuantitativa	Wikipedia
Generación de resúmenes	summarization model	modelo de sumarizacion	Hugging Face Tasks
	summarization models	modelos de sumarizacion	Hugging Face Tasks
	automatic summarization	resumen automatico	Wikipedia - Common NLP tasks
Análisis sintáctico	sentence boundary detection	deteccion de limimtes de oraciones	
		deteccion de limites de frases	(Bikel and Zitouni, 2012)
	grammar induction	induccion gramatical	Wikipedia - Common NLP tasks
	grammar inference	inferencia gramatical	Wikipedia - Common NLP tasks
	sentence boundary disambiguation	desambiguacion de limites de oraciones	
		desambiguacion de limites de frases	Wikipedia - Common NLP tasks
	syntactic parsing	parseo sintactico	Wikipedia - Common NLP tasks
Análisis de texto	text classification model	modelo de clasificacion de texto	
		modelo de clasificacion de textos	Hugging Face Tasks
	text classification models	modelos de clasificacion de texto	
		modelos de clasificacion de textos	Hugging Face Tasks
	automatic text analytics	analisis de texto automatico	Wikipedia
	text mining	mineria de texto	Wikipedia
Procesamiento de voz	speech processing	procesamiento de habla	
		procesamiento del habla	Wikipedia - Common NLP tasks
	speech recognition	reconocimiento de habla	
		reconocimiento del habla	Wikipedia - Common NLP tasks
	speech segmenatation	segmentacion de habla	
		segmentacion del habla	Wikipedia - Common NLP tasks
	text to speech	de texto a voz	Wikipedia - Common NLP tasks
Traducción	translation model	modelo de traduccion	Hugging Face Tasks
	translation models	modelos de traduccion	Hugging Face Tasks
	automatic translation	traducción automatica	Wikipedia - Common NLP tasks
	machine translation		Wikipedia - Common NLP tasks

B. Bolsa de expresiones regulares para la detección de atributos

ATRIBUTO	INGLÉS	ESPAÑOL
Rendimiento		precis[oa]
		correct[oa]
		falla
		imprecis[oa]
		desempeñ[oa]
	precise	precision
	accurate	equivoca
	correct	error
	failure	errores
	imprecise	rendimiento
	precision	eficacia
	wrong	eficiente
	performance	eficiencia
	efficiency	efectividad
	efficient	procesador
	efficacy	recursos
	effectiveness	consum[oe]
	processor	memoria
	resources	energia
	memory	almacenamiento
	energy	ancho de banda
	storage	lent[ao]
	bandwidth	rapid[ao]
	slow	funciona (?:bien|mal|regular|peor|mejor)
	fast	
	works (?:well|bad|fine|worse|better)	
Usabilidad	security	seguridad
	privacy	privacidad
	login	login
	session	sesion
	password	contrasena
	key	password
	breach	violacion
	infringement	infraccion
	credentials	credenciales
	permission	permiso
	cookies	cookies
	insecure	acceder
	-	insegur[oa]
Precio	cheap	car[oa]
	cheaper	carisim[oa]
	expensive	barat[oa]
	free of charge	baratisim[oa]
	(?:completely|totally|for) free	gratis
	gratuitous	economic[oa]
	economical	de pago
	paid	licencia
	license	abonar
	subscribe	suscribir
	suscripcion	suscripcion

C. Diseño de encuestas

Cuestionario adopción y satisfacción

Notas para scripter en COLOR AMARILLO.

CLIENTE:

MEDICION: 2023

STAKEHOLDER: PÚBLICO GENERAL

SECCIÓN A. INTRODUCCIÓN

BIENVENIDA

SECCIÓN B. DATOS DEL ENTREVISTADO

Edad. ¿Cuál es tu edad?

EXCLUIR MENORES DE 18 Y MAYORES DE 85

Género. ¿Cuál es tu género?

Masculino	
Femenino	
Prefiero no contestar	

CCAA NO SE PREGUNTA VIENE DIRECTAMENTE DE LA BASE DE DATOS DEL PANEL PUEDEN PARTICIPAR DE TODAS LAS CCAA, DISTRIBUCION ESPERADA PROPORCIONAL A LA POBLACION

NSE NO SE PREGUNTA, INCLUIR TODOS LOS SEGMENTOS DE A, B y C. DIRECTAMENTE DE LA BASE DE DATOS DEL PANEL

SECCIÓN C. ADOPCION

P1. Has utilizado alguna vez: ROTAMOS

		Si	No
P1.1	Alguna herramienta para análisis de opiniones en español		
P1.2	Algún asistente virtual en español		
P1.3	Alguna herramienta para la traducción automática de textos en español o al español		
P1.4	Alguna herramienta de teclado predictivo en español (los teclados predictivos te sugieren o corrigen palabras a medida que escribes texto en el teclado)		
P1.5	Algún buscador web en español		

ROTAR EL ORDEN DE PRESENTACION DE LAS HERRAMIENTAS ENTRE ENCUESTADOS

SÓLO CONTINUAR CON LAS SECCIONES CORRESPONDIENTES A LAS P.1 DONDE HAYA RESPONDID0O AFIRMATIVAMENTE

SI UNA PERSONA RESPONDE NO A TODAS LAS P.1, ES DECIR SI LA PERSONA NO HA UTILIZADO NINGUNA HERRAMIENTA, FINALIZA LA ENCUESTA (CREEMOS QUE ESTO NO VA A SUCEDER EN EL PANEL ON LINE, PERO HAY QUE PREVER LA PROGRAMACION DE ESTA FINALIZACION)

SECCIÓN D. ANALISIS DE OPINIONES

P2. ¿Has utilizado alguna vez alguna de estas soluciones para el análisis de opiniones en español?

		Sí, para uso personal	Sí, para uso profesional	Sí, para ambas	No
P2.1	Sprinklr	1	2	3	99
P2.2	Khoros	1	2	3	99
P2.3	NetBase Quid	1	2	3	99
P2.4	Brandwatch	1	2	3	99
P2.5	Linkfluence	1	2	3	99
P2.6	Synthesio	1	2	3	99
P2.7	Talkwalker	1	2	3	99
P2.8	Digimind	1	2	3	99
P2.9	Resonate	1	2	3	99
P2.10	Sysomos	1	2	3	99

ROTAR EL ORDEN DE PRESENTACION DE LAS HERRAMIENTAS ENTRE ENCUESTADOS

LAS HERRAMIENTAS QUE EN P2 TIENEN CODIGO 1, 2 o 3 PASAN A P3, LAS QUE TIENEN CODIGO 4 SE OMITEN, SI EN TODAS HAY CODIGO 4 PASAR A PROXIMA SECCION

P3. ¿Cuál es el grado de satisfacción con estas soluciones de análisis de opiniones en español?

		★	★ ★	★ ★ ★	★ ★ ★ ★	★ ★ ★ ★ ★
P3.1	Sprinklr	1	2	3	4	5
P3.2	Khoros	1	2	3	4	5
P3.3	NetBase Quid	1	2	3	4	5
P3.4	Brandwatch	1	2	3	4	5
P3.5	Linkfluence	1	2	3	4	5
P3.6	Synthesio	1	2	3	4	5
P3.7	Talkwalker	1	2	3	4	5
P3.8	Digimind	1	2	3	4	5
P3.9	Resonate	1	2	3	4	5
P3.10	Sysomos	1	2	3	4	5

SE PUEDE MANTENER LA ROTACION USADA EN LA PREGUNTA ANTERIOR, YA AQUÍ LA ROTACIÓN NO ES IMPORTANTE

P4. ¿Has observado alguna de estas limitaciones a la hora de utilizar estas herramientas en español? Si no ha observado limitaciones, marque “ninguna limitación”.

TABLA CON LAS HERRAMIENTAS QUE EN P2 TIENEN CODIGO 1, 2 o 3 SI SE NECESITAN 2 O MAS HOJAS PORQUE EVALUE MUCHAS HERRAMIENTAS SE PUEDE HACER

		Herramienta x	Herramienta x	Herramienta x	Herramienta x
1	En el desempeño o rendimiento				
2	En las funcionalidades				
3	En la compatibilidad con otros sistemas o con el equipo o dispositivo				
4	De seguridad				
5	De privacidad				
6	De precio				
7	En la comprensión del español				
	Aporta información sesgada				
	Aporta información engañosa				
	Solo aporta información relativamente obvia				
98	Otras limitaciones				
99	Ninguna limitación				

EL ENCUESTADO PUEDE MARCAR VARIAS LIMITACIONES POR HERRAMIENTA, PERO SI USA ULTIMA OPCION DE NINGUNA, DEBE SER UNICA Y NO PODRÁ SELECCIONAR OTRA

ROTAR TODAS MENOS 98 Y 99

SECCIÓN E. ASISTENTES VIRTUALES

P5. ¿Has utilizado alguna vez alguna de estas soluciones de asistencia virtual en español?

		Sí, para uso personal	Sí, para uso profesional	Sí, para ambas	No
P5.1	Google Assistant	1	2	3	99
P5.2	Siri	1	2	3	99
P5.3	Alexa	1	2	3	99
P5.4	Bixby	1	2	3	99
~~P5.5~~	~~Cortana~~	~~1~~	~~2~~	~~3~~	~~99~~
P5.6	Kore.ai	1	2	3	99
P5.7	IBM Watson Assistant	1	2	3	99
P5.8	Amazon Lex	1	2	3	99

P5.9	Google Dialogflow	1	2	3	99
P5.10	Amelia	1	2	3	99
P5.11	ChatGPT	1	2	3	99
	Google Bard				

ROTAR EL ORDEN DE PRESENTACION DE LAS HERRAMIENTAS ENTRE ENCUESTADOS

LAS HERRAMIENTAS QUE EN P5 TIENEN CODIGO 1, 2 o 3 PASAN A P6, LAS QUE TIENEN CODIGO 4 SE OMITEN, SI EN TODAS HAY CODIGO 4 PASAR A PROXIMA SECCION

P6. ¿Cuál es el grado de satisfacción con estos asistentes virtuales en español?

		★	★ ★	★ ★ ★	★ ★ ★ ★	★ ★ ★ ★ ★
P6.1	Google Assistant	1	2	3	4	5
P6.2	Siri	1	2	3	4	5
P6.3	Alexa	1	2	3	4	5
P6.4	Bixby	1	2	3	4	5
~~P6.5~~	~~Cortana~~	~~1~~	~~2~~	~~3~~	~~4~~	~~5~~
P6.6	Kore.ai	1	2	3	4	5
P6.7	IBM Watson Assistant	1	2	3	4	5
P6.8	Amazon Lex	1	2	3	4	5
P6.9	Google Dialogflow	1	2	3	4	5
P6.10	Amelia	1	2	3	4	5
P6.11	ChatGPT	1	2	3	4	5
	Google Bard					

SE PUEDE MANTENER LA ROTACION USADA EN LA PREGUNTA ANTERIOR, YA QUE AQUÍ LA ROTACIÓN NO ES IMPORTANTE

P7. ¿Has observado alguna de estas limitaciones a la hora de utilizar estos asistentes en español? Si no ha observado limitaciones marque en “ninguna limitación”.

TABLA CON LAS HERRAMIENTAS QUE EN P5 TIENEN CODIGO 1, 2 o 3 SI SE NECESITAN 2 O MAS HOJAS PORQUE EVALUE MUCHOS ASISTENTES SE PUEDE HACER

		Herramienta x	Herramienta x	Herramienta x	Herramienta x
1	En el desempeño o rendimiento				
2	En las funcionalidades				
3	En la compatibilidad con otros sistemas o con el				

	equipo o dispositivo				
4	De seguridad				
5	De privacidad				
6	De precio				
7	En la comprensión del español				
	Proporciona contenidos ofensivos, tóxicos o inadecuados				
	Proporciona contenidos sesgados y/o estereotipados				
	Proporciona información convincente a primera vista, pero errónea al verificarla				
	Sólo ofrece información genérica que ya se conocía				
	No justifica las respuestas ni da acceso a las fuentes que la justifican				
98	Otras limitaciones				
99	Ninguna limitación				

EL ENCUESTADO PUEDE MARCAR VARIAS LIMITACIONES POR HERRAMIENTA, PERO SI USA ULTIMA OPCION DE NINGUNA, DEBE SER UNICA Y NO PODRÁ SELECCIONAR OTRA

ROTAR TODAS MENOS 98 Y 99

SECCIÓN F. TRADUCCIÓN AUTOMATICA

P8. ¿Has utilizado alguna vez alguna de estas soluciones para la traducción automática de textos en o al español?

		Sí, para uso personal	Sí, para uso profesional	Sí, para ambas	No
P8.1	Google Translate	1	2	3	99
P8.2	DeepL	1	2	3	99
P8.3	Bing Translator o Microsoft Translator	1	2	3	99
P8.4	Amazon Translate	1	2	3	99
P8.5	Systran Translate	1	2	3	99
P8.6	Reverso Translator	1	2	3	99
P8.7	memoQ Translator PRO	1	2	3	99
P8.8	Smartling	1	2	3	99
P8.9	Crowdin	1	2	3	99

P8.10	TextUnited	1	2	3	99
	ChatGPT				
	Google Bard				

ROTAR EL ORDEN DE PRESENTACION DE LAS HERRAMIENTAS ENTRE ENCUESTADOS

LAS HERRAMIENTAS QUE EN P8 TIENEN CODIGO 1, 2 o 3 PASAN A P9, LAS QUE TIENEN CODIGO 4 SE OMITEN, SI EN TODAS HAY CODIGO 4 PASAR A PROXIMA SECCION

P9. ¿Cuál es el grado de satisfacción con estas soluciones de traducción automática en o al español?

		★	★ ★	★ ★ ★	★ ★ ★ ★	★ ★ ★ ★ ★
P9.1	Google Translate	1	2	3	4	5
P9.2	DeepL	1	2	3	4	5
P9.3	Bing Translator o Microsoft Translator	1	2	3	4	5
P9.4	Amazon Translate	1	2	3	4	5
P9.5	Systran Translate	1	2	3	4	5
P9.6	Reverso Translator	1	2	3	4	5
P9.7	memoQ Translator PRO	1	2	3	4	5
P9.8	Smartling	1	2	3	4	5
P9.9	Crowdin	1	2	3	4	5
P9.10	TextUnited	1	2	3	4	5
	ChatGPT					
	Google Bard					

SE PUEDE MANTENER LA ROTACION USADA EN LA PREGUNTA ANTERIOR, YA QUE AQUÍ LA ROTACIÓN NO ES IMPORTANTE

P10. ¿Has observado alguna de estas limitaciones a la hora de utilizar estas herramientas de traducción automática en o al español? Si no ha observado limitaciones marque en “ninguna limitación”.

TABLA CON LAS HERRAMIENTAS QUE EN P8 TIENEN CODIGO 1, 2 o 3 SI SE NECESITAN 2 O MAS HOJAS PORQUE EVALUE MUCHAS HERRAMIENTAS SE PUEDE HACER

		Herramienta	Herramienta	Herramienta	Herramienta

		x	x	x	x
1	En el desempeño o rendimiento				
2	En las funcionalidades				
3	En la compatibilidad con otros sistemas o con el equipo o dispositivo				
4	De seguridad				
5	De privacidad				
6	De precio				
7	En un ámbito o sector específico (por ejemplo, traducción jurídica)				
	Presenta sesgos sistemáticos de traducción y/o estereotipados				
	Proporciona traducciones naturales y fluidas pero incorrectas				
	Traducciones literales que no captura el trasfondo del texto				
98	Otras limitaciones				
99	Ninguna limitación				

EL ENCUESTADO PUEDE MARCAR VARIAS LIMITACIONES POR HERRAMIENTA, PERO SI USA ULTIMA OPCION DE NINGUNA, DEBE SER UNICA Y NO PODRÁ SELECCIONAR OTRA

ROTAR TODAS MENOS 98 Y 99

SECCIÓN G. TECLADOS PREDICTIVOS

P11. ¿Has utilizado alguna vez alguno de estos teclados predictivos en español?

		Sí, para uso personal	Sí, para uso profesional	Sí, para ambas	No
P11.1	Microsoft SwiftKey	1	2	3	99
P11.2	GBoard	1	2	3	99
P11.3	Grammarly (funcionalidad de predicción de frases)	1	2	3	99
P11.4	Fleksy	1	2	3	99

P11.5	iPhone (teclado por defecto)	1	2	3	99
~~P11.6~~	~~Phraseboard~~	~~1~~	~~2~~	~~3~~	~~99~~
P11.7	GMail (funciones predictivas en la redacción)	1	2	3	99
P11.8	Google Workspaces (funciones predictivas en la redacción)	1	2	3	99
P11.9	Microsoft Outlook (funciones predictivas en la redacción)	1	2	3	99
P11.10	Microsoft Office 365 (funciones predictivas en la redacción)	1	2	3	99

ROTAR EL ORDEN DE PRESENTACION DE LAS HERRAMIENTAS ENTRE ENCUESTADOS

LAS HERRAMIENTAS QUE EN P11 TIENEN CODIGO 1, 2 o 3 PASAN A P9, LAS QUE TIENEN CODIGO 4 SE OMITEN, SI EN TODAS HAY CODIGO 4 PASAR A PROXIMA SECCION

P12. ¿Cuál es el grado de satisfacción con estos teclados predictivos en español?

		★	★ ★	★ ★ ★	★ ★ ★ ★	★ ★ ★ ★ ★
P12.1	Microsoft SwiftKey	1	2	3	4	5
P12.2	GBoard	1	2	3	4	5
P12.3	Grammarly (funcionalidad de predicción de frases)	1	2	3	4	5
P12.4	Fleksy	1	2	3	4	5
P12.5	iPhone (teclado por defecto)	1	2	3	4	5
~~P12.6~~	~~Phraseboard~~	~~1~~	~~2~~	~~3~~	~~4~~	~~5~~
P12.7	GMail (funciones predictivas en la redacción)	1	2	3	4	5
P12.8	Google Workspaces (funciones predictivas en la redacción)	1	2	3	4	5
P12.9	Microsoft	1	2	3	4	5

	Outlook (funciones predictivas en la redacción)					
P12.10	Microsoft Office 365 (funciones predictivas en la redacción)	1	2	3	4	5

SE PUEDE MANTENER LA ROTACION USADA EN LA PREGUNTA ANTERIOR, YA QUE AQUÍ LA ROTACIÓN NO ES IMPORTANTE

P13. ¿Has observado alguna de estas limitaciones a la hora de utilizar estos teclados predictivos en español? Si no ha observado limitaciones marque en "ninguna limitación".

TABLA CON LAS HERRAMIENTAS QUE EN P11 TIENEN CODIGO 1, 2 o 3 SI SE NECESITAN 2 O MAS HOJAS PORQUE EVALUE MUCHAS HERRAMIENTAS SE PUEDE HACER

		Herramienta x	Herramienta x	Herramienta x	Herramienta x
1	En el desempeño o rendimiento				
2	En las funcionalidades				
3	En la compatibilidad con otros sistemas o con el equipo o dispositivo				
4	De seguridad				
5	De privacidad				
6	De precio				
	Los términos que sugiere están sesgados y/o responden a estereotipos				
	Sugiere palabras que "suenan bien", pero no son las más apropiadas				
	Lo que sugiere es previsible y convencional				
98	Otras limitaciones				
99	Ninguna limitación				

EL ENCUESTADO PUEDE MARCAR VARIAS LIMITACIONES POR HERRAMIENTA, PERO SI USA ULTIMA OPCION DE NINGUNA, DEBE SER UNICA Y NO PODRÁ SELECCIONAR OTRA

ROTAR TODAS MENOS 98 Y 99

SECCIÓN H. BUSCADORES WEB

P14. ¿Has utilizado alguna vez alguno de estos buscadores web en español?

		Sí, para uso personal	Sí, para uso profesional	Sí, para ambas	No
P14.1	Google	1	2	3	99
P14.2	Bing	1	2	3	99
P14.3	Yahoo Search	1	2	3	99
P14.4	DuckDuckGo	1	2	3	99
P14.5	Brave Search	1	2	3	99
P14.6	Elastic	1	2	3	99
P14.7	Mindbreeze	1	2	3	99
P.14.8	Apache Solr	1	2	3	99
	Perplexity				

ROTAR EL ORDEN DE PRESENTACION DE LAS HERRAMIENTAS ENTRE ENCUESTADOS

LAS HERRAMIENTAS QUE EN P14 TIENEN CODIGO 1, 2 o 3 PASAN A P9, LAS QUE TIENEN CODIGO 4 SE OMITEN, SI EN TODAS HAY CODIGO 4 PASAR A PROXIMA SECCION

P15. ¿Cuál es el grado de satisfacción con estos buscadores web en español?

		★	★ ★	★ ★ ★	★ ★ ★ ★	★ ★ ★ ★ ★
P15.1	Google	1	2	3	4	5
P15.2	Bing	1	2	3	4	5
P15.3	Yahoo Search	1	2	3	4	5
P15.4	DuckDuckGo	1	2	3	4	5
P15.5	Brave Search	1	2	3	4	5
P15.6	Elastic	1	2	3	4	5
P15.7	Mindbreeze	1	2	3	4	5
P15.8	Apache Solr	1	2	3	4	5
	Perplexity					

SE PUEDE MANTENER LA ROTACION USADA EN LA PREGUNTA ANTERIOR, YA AQUÍ LA ROTACIÓN NO ES IMPORTANTE

P16. ¿Has observado alguna de estas limitaciones a la hora de utilizar estos buscadores web en español? Si no ha observado limitaciones marque en “ninguna limitación”.

TABLA CON LAS HERRAMIENTAS QUE EN P11 TIENEN CODIGO 1, 2 o 3 SI SE NECESITAN 2 O MAS HOJAS PORQUE EVALUE MUCHAS HERRAMIENTAS SE PUEDE HACER

		Herramienta x	Herramienta x	Herramienta x	Herramienta x
1	En el desempeño o rendimiento				
2	En las funcionalidades				
3	En la compatibilidad con otros sistemas o con el equipo o dispositivo				
4	De seguridad				
5	De privacidad				
6	De precio				
	Devuelve páginas con contenidos tóxicos, agresivos o inapropiados en general				
	Sesgos / preferencias sistemáticas hacia determinados tipos de páginas				
	Devuelve páginas relacionadas con la consulta pero que contienen información engañosa				
	Devuelve páginas con información previsible con poco valor añadido				
98	Otras limitaciones				
99	Ninguna limitación				

EL ENCUESTADO PUEDE MARCAR VARIAS LIMITACIONES POR HERRAMIENTA, PERO SI USA ULTIMA OPCION DE NINGUNA, DEBE SER UNICA Y NO PODRÁ SELECCIONAR OTRA

ROTAR TODAS MENOS 98 Y 99

Cuestionario adopción y satisfacción

NO TRADUCIR: Notas para scripter en COLOR AMARILLO.

Age. How old are you?

EXCLUIR MENORES DE 18 Y MAYORES DE 85

Gender. What is your gender?

Male	
Female	
I rather not answer	

CCAA NO SE PREGUNTA VIENE DIRECTAMENTE DE LA BASE DE DATOS DEL PANEL PUEDEN PARTICIPAR DE TODAS LAS CCAA, DISTRIBUCION ESPERADA PROPORCIONAL A LA POBLACION

NSE NO SE PREGUNTA, INCLUIR TODOS LOS SEGMENTOS DE A, B y C. DIRECTAMENTE DE LA BASE DE DATOS DEL PANEL

C. ADOPTION

P1. Have you ever used: ROTAMOS

		Yes	No
P1.1	Any tool for opinion analysis in English		
P1.2	Any virtual assistant in English		

P1.3	Any tool for automatic translation of texts in English or into English		
P1.4	Any predictive keyboard tool in English (predictive keyboards suggest or correct words as you type text on the keyboard).		
P1.5	Any web search engine in English		

ROTAR EL ORDEN DE PRESENTACION DE LAS HERRAMIENTAS ENTRE ENCUESTADOS

SÓLO CONTINUAR CON LAS SECCIONES CORRESPONDIENTES A LAS P.1 DONDE HAYA RESPONDID0O AFIRMATIVAMENTE

SI UNA PERSONA RESPONDE NO A TODAS LAS P.1, ES DECIR SI LA PERSONA NO HA UTILIZADO NINGUNA HERRAMIENTA, FINALIZA LA ENCUESTA (CREEMOS QUE ESTO NO VA A SUCEDER EN EL PANEL ON LINE, PERO HAY QUE PREVER LA PROGRAMACION DE ESTA FINALIZACION)

SECCIÓN D. OPINION ANALYSIS

P2. ¿Has utilizado alguna vez alguna de estas soluciones para el análisis de opiniones en inglés?

		Yes, for personal use	Yes, for professional use	Yes, for both	No
P2.1	Sprinklr	1	2	3	99
P2.2	Khoros	1	2	3	99
P2.3	NetBase Quid	1	2	3	99
P2.4	Brandwatch	1	2	3	99
P2.5	Linkfluence	1	2	3	99
P2.6	Synthesio	1	2	3	99
P2.7	Talkwalker	1	2	3	99
P2.8	Digimind	1	2	3	99
P2.9	Resonate	1	2	3	99
P2.10	Sysomos	1	2	3	99

ROTAR EL ORDEN DE PRESENTACION DE LAS HERRAMIENTAS ENTRE ENCUESTADOS

LAS HERRAMIENTAS QUE EN P2 TIENEN CODIGO 1, 2 o 3 PASAN A P3, LAS QUE TIENEN CODIGO 4 SE OMITEN, SI EN TODAS HAY CODIGO 4 PASAR A PROXIMA SECCION

P3. How satisfied are you with these opinion analysis solutions in English?

		★	★ ★	★ ★ ★	★ ★ ★ ★	★ ★ ★ ★ ★
P3.1	Sprinklr	1	2	3	4	5
P3.2	Khoros	1	2	3	4	5
P3.3	NetBase Quid	1	2	3	4	5
P3.4	Brandwatch	1	2	3	4	5
P3.5	Linkfluence	1	2	3	4	5
P3.6	Synthesio	1	2	3	4	5
P3.7	Talkwalker	1	2	3	4	5
P3.8	Digimind	1	2	3	4	5
P3.9	Resonate	1	2	3	4	5
P3.10	Sysomos	1	2	3	4	5

SE PUEDE MANTENER LA ROTACION USADA EN LA PREGUNTA ANTERIOR, YA AQUÍ LA ROTACIÓN NO ES IMPORTANTE

P4. Have you observed any of these limitations when using these tools? If you have not observed any limitations, check "no limitations".

TABLA CON LAS HERRAMIENTAS QUE EN P2 TIENEN CODIGO 1, 2 o 3 SI SE NECESITAN 2 O MAS HOJAS PORQUE EVALUE MUCHAS HERRAMIENTAS SE PUEDE HACER

		Tool x	Tool x	Tool x	Tool x
1	In performance				
2	In functionalities				
3	On compatibility with other systems or with the equipment or device				
4	In security				
5	In privacy				
6	In price				
7	In understanding English				
	Provides biased information				
	Provides misleading information				
	Only provides relatively obvious information				
98	Other limitations				
99	No limitations				

EL ENCUESTADO PUEDE MARCAR VARIAS LIMITACIONES POR HERRAMIENTA, PERO SI USA ULTIMA OPCION DE NINGUNA, DEBE SER UNICA Y NO PODRÁ SELECCIONAR OTRA

SECCIÓN E. VIRTUAL ASSISTANTS

P5. Have you ever used any of these virtual assistants in English?

		Yes, for personal use	Yes, for professional use	Yes, for both	No
P5.1	Google Assistant	1	2	3	99
P5.2	Siri	1	2	3	99
P5.3	Alexa	1	2	3	99
P5.4	Bixby	1	2	3	99
~~P5.5~~	~~Cortana~~	~~1~~	~~2~~	~~3~~	~~99~~
P5.6	Kore.ai	1	2	3	99
P5.7	IBM Watson Assistant	1	2	3	99
P5.8	Amazon Lex	1	2	3	99
P5.9	Google Dialogflow	1	2	3	99
P5.10	Amelia	1	2	3	99
P5.11	ChatGPT	1	2	3	99

	Google Bard					

ROTAR EL ORDEN DE PRESENTACION DE LAS HERRAMIENTAS ENTRE ENCUESTADOS

LAS HERRAMIENTAS QUE EN P5 TIENEN CODIGO 1, 2 o 3 PASAN A P6, LAS QUE TIENEN CODIGO 4 SE OMITEN, SI EN TODAS HAY CODIGO 4 PASAR A PROXIMA SECCION

P6. How satisfied are you with these virtual assistants in English?

		★	★ ★	★ ★ ★	★ ★ ★ ★	★ ★ ★ ★ ★
P6.1	Google Assistant	1	2	3	4	5
P6.2	Siri	1	2	3	4	5
P6.3	Alexa	1	2	3	4	5
P6.4	Bixby	1	2	3	4	5
~~P6.5~~	~~Cortana~~	~~1~~	~~2~~	~~3~~	~~4~~	~~5~~
P6.6	Kore.ai	1	2	3	4	5
P6.7	IBM Watson Assistant	1	2	3	4	5
P6.8	Amazon Lex	1	2	3	4	5
P6.9	Google Dialogflow	1	2	3	4	5
P6.10	Amelia	1	2	3	4	5
P6.11	ChatGPT	1	2	3	4	5
	Google Bard					

SE PUEDE MANTENER LA ROTACION USADA EN LA PREGUNTA ANTERIOR, YA QUE AQUÍ LA ROTACIÓN NO ES IMPORTANTE

P7. Have you observed any of these limitations when using these virtual assistants in English? If you have not observed any limitations, please check "no limitations".

TABLA CON LAS HERRAMIENTAS QUE EN P5 TIENEN CODIGO 1, 2 o 3 SI SE NECESITAN 2 O MAS HOJAS PORQUE EVALUE MUCHOS ASISTENTES SE PUEDE HACER

		Tool x	Tool x	Tool x	Tool x
1	In performance				
2	In functionalities				
3	On compatibility with other systems or with the equipment or device				
4	In security				
5	In privacy				
6	In price				
7	In understanding English				

	Provides offensive, toxic or inappropriate content				
	Provides biased and/or stereotyped content				
	Provides information that is convincing at first glance, but incorrect upon verification				
	Only offers generic information that was already known				
	Does not justify answers or provide access to the sources that justify them				
98	Other limitations				
99	No limitations				

EL ENCUESTADO PUEDE MARCAR VARIAS LIMITACIONES POR HERRAMIENTA, PERO SI USA ULTIMA OPCION DE NINGUNA, DEBE SER UNICA Y NO PODRÁ SELECCIONAR OTRA

SECCIÓN F. AUTOMATIC TRANSLATION

P8. Have you ever used any of these solutions for the automatic translation of texts in or into English?

		Yes, for personal use	Yes, for professional use	Yes, for both	No
P8.1	Google Translate	1	2	3	99
P8.2	DeepL	1	2	3	99
P8.3	Bing Translator o Microsoft Translator	1	2	3	99
P8.4	Amazon Translate	1	2	3	99
P8.5	Systran Translate	1	2	3	99
P8.6	Reverso Translator	1	2	3	99
P8.7	memoQ Translator PRO	1	2	3	99
P8.8	Smartling	1	2	3	99
P8.9	Crowdin	1	2	3	99
P8.10	TextUnited	1	2	3	99
	ChatGPT				
	Google Bard				

ROTAR EL ORDEN DE PRESENTACION DE LAS HERRAMIENTAS ENTRE ENCUESTADOS

LAS HERRAMIENTAS QUE EN P8 TIENEN CODIGO 1, 2 o 3 PASAN A P9, LAS QUE TIENEN CODIGO 4 SE OMITEN, SI EN TODAS HAY CODIGO 4 PASAR A PROXIMA SECCION

P9. How satisfied are you with these solutions for the automatic translation of texts into English?

		★	★ ★	★ ★ ★	★ ★ ★ ★	★ ★ ★ ★ ★
P9.1	Google Translate	1	2	3	4	5
P9.2	DeepL	1	2	3	4	5
P9.3	Bing Translator o Microsoft Translator	1	2	3	4	5
P9.4	Amazon Translate	1	2	3	4	5
P9.5	Systran Translate	1	2	3	4	5
P9.6	Reverso Translator	1	2	3	4	5
P9.7	memoQ Translator PRO	1	2	3	4	5
P9.8	Smartling	1	2	3	4	5
P9.9	Crowdin	1	2	3	4	5
P9.10	TextUnited	1	2	3	4	5
	ChatGPT					
	Google Bard					

SE PUEDE MANTENER LA ROTACION USADA EN LA PREGUNTA ANTERIOR, YA QUE AQUÍ LA ROTACIÓN NO ES IMPORTANTE

P10. Have you observed any of these limitations when using these tools for automatic translation in or into English? If you have not observed any limitations, please check "no limitations".

TABLA CON LAS HERRAMIENTAS QUE EN P8 TIENEN CODIGO 1, 2 o 3 SI SE NECESITAN 2 O MAS HOJAS PORQUE EVALUE MUCHAS HERRAMIENTAS SE PUEDE HACER

		Tool x	Tool x	Tool x	Tool x
1	In performance				
2	In functionalities				
3	On compatibility with other systems or with the equipment or device				
4	In security				
5	In privacy				

6	In price				
7	In a specific sector or area (for instance, legal texts translation)				
	Presents systematic and/or stereotyped translation biases				
	Provides natural and fluent translations but incorrect ones				
	Literal translations that do not capture the background of the text				
98	Other limitations				
99	No limitations				

EL ENCUESTADO PUEDE MARCAR VARIAS LIMITACIONES POR HERRAMIENTA, PERO SI USA ULTIMA OPCION DE NINGUNA, DEBE SER UNICA Y NO PODRÁ SELECCIONAR OTRA

SECCIÓN G. PREDICTIVE KEYBOARDS

P11. ¿Has utilizado alguna vez alguno de estos teclados predictivos en inglés?

		Yes, for personal use	Yes, for professional use	Yes, for both	No
P11.1	Microsoft SwiftKey	1	2	3	99
P11.2	GBoard	1	2	3	99
P11.3	Grammarly (phrasal predictions feature)	1	2	3	99
P11.4	Fleksy	1	2	3	99
P11.5	iPhone's default keyboard	1	2	3	99
~~P11.6~~	~~Phraseboard~~	~~1~~	~~2~~	~~3~~	~~99~~
P11.7	GMail (predictive functions in writing)	1	2	3	99
P11.8	Google Workspaces (predictive functions in writing)	1	2	3	99
P11.9	Microsoft Outlook (predictive functions in writing)	1	2	3	99
P11.10	Microsoft Office 365 (predictive functions in writing)	1	2	3	99

ROTAR EL ORDEN DE PRESENTACION DE LAS HERRAMIENTAS ENTRE ENCUESTADOS

LAS HERRAMIENTAS QUE EN P11 TIENEN CODIGO 1, 2 o 3 PASAN A P9, LAS QUE TIENEN CODIGO 4 SE OMITEN, SI EN TODAS HAY CODIGO 4 PASAR A PROXIMA SECCION

P12. How satisfied are you with these predictive keyboards in English?

		★	★ ★	★ ★ ★	★ ★ ★ ★	★ ★ ★ ★ ★
P12.1	Microsoft SwiftKey	1	2	3	4	5
P12.2	GBoard	1	2	3	4	5
P12.3	Grammarly (phrasal predictions feature)	1	2	3	4	5
P12.4	Fleksy	1	2	3	4	5
P12.5	iPhone's default keyboard	1	2	3	4	5
~~P12.6~~	~~Phraseboard~~	~~1~~	~~2~~	~~3~~	~~4~~	~~5~~
P12.7	GMail (predictive functions in writing)	1	2	3	4	5
P12.8	Google Workspaces (predictive functions in writing)	1	2	3	4	5
P12.9	Microsoft Outlook (predictive functions in writing)	1	2	3	4	5
P12.10	Microsoft Office 365 (predictive functions in writing)	1	2	3	4	5

SE PUEDE MANTENER LA ROTACION USADA EN LA PREGUNTA ANTERIOR, YA QUE AQUÍ LA ROTACIÓN NO ES IMPORTANTE

P13. Have you observed any of these limitations when using these predictive keyboards in English? If you have not observed any limitations, please check "no limitations".

TABLA CON LAS HERRAMIENTAS QUE EN P11 TIENEN CODIGO 1, 2 o 3 SI SE NECESITAN 2 O MAS HOJAS PORQUE EVALUE MUCHAS HERRAMIENTAS SE PUEDE HACER

		Tool x	Tool x	Tool x	Tool x
1	In performance				
2	In functionalities				
3	On compatibility with other systems or with the equipment or device				
4	In security				
5	In privacy				
6	In price				
	The terms it suggests are biased and/or respond to stereotypes				
	Suggests words that "sound good", but are not the most appropriate				
	What it suggests is predictable and conventional				
98	Other limitations				
99	No limitations				

EL ENCUESTADO PUEDE MARCAR VARIAS LIMITACIONES POR HERRAMIENTA, PERO SI USA ULTIMA OPCION DE NINGUNA, DEBE SER UNICA Y NO PODRÁ SELECCIONAR OTRA

SECCIÓN H. WEB SEARCH ENGINES

P14. Have you ever used any of these web search engines in English?

		Yes, for personal use	Yes, for professional use	Yes, for both	No
P14.1	Google	1	2	3	99
P14.2	Bing	1	2	3	99
P14.3	Yahoo Search	1	2	3	99
P14.4	DuckDuckGo	1	2	3	99
P14.5	Brave Search	1	2	3	99
P14.6	Elastic	1	2	3	99
P14.7	Mindbreeze	1	2	3	99
P.14.8	Apache Solr	1	2	3	99
	Perplexity				

ROTAR EL ORDEN DE PRESENTACION DE LAS HERRAMIENTAS ENTRE ENCUESTADOS

LAS HERRAMIENTAS QUE EN P14 TIENEN CODIGO 1, 2 o 3 PASAN A P9, LAS QUE TIENEN CODIGO 4 SE OMITEN, SI EN TODAS HAY CODIGO 4 PASAR A PROXIMA SECCION

P15. How satisfied are you with these web search engines in English?

		★	★ ★	★ ★ ★	★ ★ ★ ★	★ ★ ★ ★ ★
P15.1	Google	1	2	3	4	5
P15.2	Bing	1	2	3	4	5
P15.3	Yahoo Search	1	2	3	4	5
P15.4	DuckDuckGo	1	2	3	4	5
P15.5	Brave Search	1	2	3	4	5
P15.6	Elastic	1	2	3	4	5
P15.7	Mindbreeze	1	2	3	4	5
P15.8	Apache Solr	1	2	3	4	5
	Perplexity					

SE PUEDE MANTENER LA ROTACION USADA EN LA PREGUNTA ANTERIOR, YA AQUÍ LA ROTACIÓN NO ES IMPORTANTE

P16. Have you observed any of these limitations when using these web search engines in English? If you have not observed any limitations, please check "no limitations".

TABLA CON LAS HERRAMIENTAS QUE EN P11 TIENEN CODIGO 1, 2 o 3 SI SE NECESITAN 2 O MAS HOJAS PORQUE EVALUE MUCHAS HERRAMIENTAS SE PUEDE HACER

		Tool x	Tool x	Tool x	Tool x
1	In performance				
2	In functionalities				
3	On compatibility with other systems or with the equipment or device				
4	In security				
5	In privacy				
6	In price				
	Returns pages with toxic, aggressive, or generally inappropriate content				
	Systematic biases/preferences towards certain types of pages				
	Returns pages related to the query but contain misleading information				
	Returns pages with predictable information with				

	little added value				
98	Other limitations				
99	No limitations				

EL ENCUESTADO PUEDE MARCAR VARIAS LIMITACIONES POR HERRAMIENTA, PERO SI USA ULTIMA OPCION DE NINGUNA, DEBE SER UNICA Y NO PODRÁ SELECCIONAR OTRA

D. Tablas de datos

Tabla 26: Análisis de opiniones.

	netbase-quid		brandwatch		linkfluence		synthesio		talkwalker		digimind		meltwater	
	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN
Clasificación de emociones	8	8	1	1	6	6	5	5	1	1	25	25	6	6
Clasificación de impacto reputacional	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Clasificación de sentimiento	2	2	1	1	2	2	2	2	2	2	2	2	2	2
Detección de bots	0	0	0	0	1	1	0	0	1	1	0	0	1	1
Detección de entidades	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Detección de mensajes inapropiados	1	1	1	1	0	0	1	1	1	1	1	1	0	0
Detección de motivaciones	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Detección de tema de conversación	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Posibilidad de ajustar el modelo	1	1	0	0	0	0	1	1	1	1	1	1	1	1
Posibilidad de definir clases	1	1	0	0	0	0	1	1	1	1	1	1	0	0
Sentimiento por aspectos del producto	1	1	0	0	1	1	1	1	1	1	1	1	0	1
Sentimiento por entidades	0	0	0	0	1	1	1	1	1	1	1	1	1	1

Tabla 27: Asistentes virtuales.

	google ass.		siri		alexa		bixby		chatgpt		koreai		ibm-watson-ass.		amazon-lex		google-dialogflow		amelia		google-bard	
	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN
Acentos regionales	6	12	4	9	4	5	2	3	1	1	0	0	0	0	3	4	3	6	0	0	0	0
Capacidad de escribir texto	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1
Comandos aceptados	33	33	134	134	137	200	119	179	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Otras funcionalidades	0	0	0	0	0	0	0	0	0	0	0	0	10	10	0	0	2	2	0	0	0	0
Posibilidad de añadir habilidades	0	0	4	4	1	1	1	1	0	0	34	34	1	1	0	0	1	1	1	1	7	7
Reconocimiento de voz	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE TRANSFORMACIÓN DIGITAL
red.es
Plan de Recuperación, Transformación y Resiliencia
UNED

Tabla 28: Traducción automática.

	google		deepl		bing		systran		amazon		reverso		memoq		smartling		crowdin		textunited		chatgpt		googlebard	
	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN
Adaptación a dominios	0	0	1	1	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Corrección de gramática	1	1	0	0	1	1	0	0	0	0	1	1	1	1	0	0	1	1	1	1	1	1	1	1
De imágenes de palabras a texto	1	1	0	0	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
De texto a texto	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
De texto a voz	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
De voz a texto	1	1	0	0	2	2	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
De voz a voz	1	1	1	1	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Detección de idioma	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1
Idiomas desde los que puede traducir	133	133	31	31	132	132	55	55	75	75	26	26	262	262	523	523	314	314	170	170	14	14	132	132
Personalización a dominios	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
Posibilidad de modificar traducciones	0	0	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1	0	0	0	0	0	0
Traducción de archivos	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
Traducción web	1	1	0	0	1	1	1	1	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
Variantes regionales	0	0	2	0	0	0	2	3	2	3	0	0	22	14	0	0	22	27	20	4	0	0	0	0
Versiones de la traducción	1	1	1	1	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	1	1	1	1

Tabla 29: Teclados predictivo.

	microsoft-swiftkey		gboard		grammarly		fleksy		iphone-keyboard		gmail		google-workspaces		microsoft-outlook		microsoft-office-365	
	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN
Autocompletado de palabras	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Corrección de gramática	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
De voz a texto	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1
Detección de idioma	1	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1
Generación de texto	1	1	1	1	0	2	2	2	0	0	2	2	2	2	2	2	1	1
Predicción personalizada	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Sugerencias de palabras	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE TRANSFORMACIÓN DIGITAL
red.es
Plan de Recuperación, Transformación y Resiliencia
UNED

Tabla 30: Buscadores web.

	google-search		bing		yahoo-search		duckduckgo		brave-search		elasticsearch		mindbreeze		apache-solr		perplexity	
	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN
Búsqueda de audio	0	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0
Búsqueda de imágenes	1	1	1	1	1	1	0	0	1	1	0	0	0	0	0	0	1	1
Búsqueda de información	1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0
Búsqueda de respuestas	1	1	1	1	0	0	1	1	1	1	0	0	1	1	0	0	1	1
Búsqueda de sinónimos	0	0	1	1	0	1	0	1	0	0	0	0	1	1	0	0	1	1
Búsqueda de texto en imágenes	1	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	1	1
Búsqueda de vídeos	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1
Clasificación de tema	8	8	0	0	2	3	0	0	2	2	1	1	0	0	1	1	0	0
Corrección de gramática	1	1	1	1	0	1	1	1	1	1	0	0	0	0	1	1	1	1
Detección de entidades	1	1	1	1	1	1	1	1	0	0	0	0	1	1	0	0	1	1
Detección de significado	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1

Tabla 31: Menciones en medios.

Análisis de opiniones	ES	EN
Sprinklr	537	19753
Khoros	48	4500
NetBase Quid	8	3138
Brandwatch	36	5603
Linkfluence	6	446
Synthesio	4	762
Talkwalker	32	3823
Digimind	17	2099
Resonate	0	43
Meltwater	118	9528
Asistentes virtuales	ES	EN
Google Assistant	2120	343688
Siri	2679	330276
Alexa	4882	917943
Bixby	352	29708
Kore.ai	5	7853
IBM Watson Assistant	0	2657
Amazon Lex	8	2712
Google Dialogflow	76	1404
Amelia	262	231
ChatGPT	12569	4402
Google Bard	85	26725
Traducción automática	ES	EN
Google Translate	938	21986
DeepL	236	5914
Bing Translator o Microsoft Translator	11	4154
Amazon Translate	1	2413
Systran Translate	15	2446
Reverso Translator	0	72
memoQ Translator PRO	1	228
Smartling	17	358
Crowdin	7	101
TextUnited	0	65
Teclados predictivos	ES	EN
Microsoft SwiftKey	2	2323
GBoard	102	8912
Grammarly Keyboard	0	285
Fleksy	24	727
iPhone Keyboard	6	7000
GMail	0	171
Google Workspace	0	35
Microsoft Outlook	0	10
Microsoft Office 365	3445	27
Buscadores web	ES	EN
Google Search	2233	209415
Bing	180	87851
Yahoo Search	24	4712
DuckDuckGo	170	34986
Brave Search	1	1949
Elasticsearch	108	15909
Mindbreeze	19	1054
Apache Solr	0	560
Perplexity	43	5184

Tabla 36: Limitaciones de análisis de opiniones.

Producto	**Tipo de limitaciones**	**Español**	**Inglés**

Sprinklr	En el desempeño o rendimiento	5 %	19 %
Sprinklr	En las funcionalidades	9 %	26 %
Sprinklr	En la compatibilidad con otros sistemas o con el equipo o dispositivo	9 %	16 %
Sprinklr	De seguridad	27 %	18 %
Sprinklr	De privacidad	23 %	20 %
Sprinklr	De precio	18 %	22 %
Sprinklr	En la comprensión del idioma	9 %	23 %
Sprinklr	Aporta información sesgada	9 %	8 %
Sprinklr	Aporta información engañosa	9 %	15 %
Sprinklr	Solo aporta información relativamente obvia	18 %	14 %
Sprinklr	Otras limitaciones	5 %	5 %
Khoros	En el desempeño o rendimiento	10 %	14 %
Khoros	En las funcionalidades	17 %	17 %
Khoros	En la compatibilidad con otros sistemas o con el equipo o dispositivo	13 %	18 %
Khoros	De seguridad	10 %	16 %
Khoros	De privacidad	17 %	11 %
Khoros	De precio	10 %	17 %
Khoros	En la comprensión del idioma	10 %	15 %
Khoros	Aporta información sesgada	20 %	13 %
Khoros	Aporta información engañosa	7 %	16 %
Khoros	Solo aporta información relativamente obvia	13 %	15 %
Khoros	Otras limitaciones	0 %	10 %
NetBase Quid	En el desempeño o rendimiento	5 %	17 %
NetBase Quid	En las funcionalidades	26 %	19 %
NetBase Quid	En la compatibilidad con otros sistemas o con el equipo o dispositivo	5 %	14 %
NetBase Quid	De seguridad	16 %	17 %
NetBase Quid	De privacidad	11 %	17 %
NetBase Quid	De precio	5 %	18 %
NetBase Quid	En la comprensión del idioma	37 %	20 %
NetBase Quid	Aporta información sesgada	11 %	10 %
NetBase Quid	Aporta información engañosa	16 %	17 %
NetBase Quid	Solo aporta información relativamente obvia	16 %	13 %
NetBase Quid	Otras limitaciones	11 %	4 %
Brandwatch	En el desempeño o rendimiento	5 %	20 %
Brandwatch	En las funcionalidades	23 %	19 %
Brandwatch	En la compatibilidad con otros sistemas o con el equipo o dispositivo	5 %	21 %
Brandwatch	De seguridad	18 %	13 %
Brandwatch	De privacidad	14 %	19 %
Brandwatch	De precio	23 %	25 %
Brandwatch	En la comprensión del idioma	5 %	21 %
Brandwatch	Aporta información sesgada	9 %	12 %
Brandwatch	Aporta información engañosa	0 %	12 %
Brandwatch	Solo aporta información relativamente obvia	5 %	16 %
Brandwatch	Otras limitaciones	5 %	6 %
Linkfluence	En el desempeño o rendimiento	11 %	16 %
Linkfluence	En las funcionalidades	7 %	14 %
Linkfluence	En la compatibilidad con otros sistemas o con el equipo o dispositivo	7 %	15 %
Linkfluence	De seguridad	11 %	19 %
Linkfluence	De privacidad	18 %	18 %
Linkfluence	De precio	18 %	18 %
Linkfluence	En la comprensión del idioma	25 %	20 %
Linkfluence	Aporta información sesgada	7 %	11 %
Linkfluence	Aporta información engañosa	4 %	17 %
Linkfluence	Solo aporta información relativamente obvia	7 %	26 %
Linkfluence	Otras limitaciones	4 %	8 %
Synthesio	En el desempeño o rendimiento	16 %	14 %
Synthesio	En las funcionalidades	16 %	20 %
Synthesio	En la compatibilidad con otros sistemas o con el equipo o dispositivo	12 %	15 %
Synthesio	De seguridad	0 %	20 %
Synthesio	De privacidad	24 %	15 %
Synthesio	De precio	8 %	14 %
Synthesio	En la comprensión del idioma	20 %	20 %
Synthesio	Aporta información sesgada	20 %	21 %
Synthesio	Aporta información engañosa	20 %	14 %
Synthesio	Solo aporta información relativamente obvia	12 %	17 %

Synthesio	Otras limitaciones	0 %	9 %
Talkwalker	En el desempeño o rendimiento	3 %	14 %
Talkwalker	En las funcionalidades	9 %	24 %
Talkwalker	En la compatibilidad con otros sistemas o con el equipo o dispositivo	6 %	18 %
Talkwalker	De seguridad	9 %	21 %
Talkwalker	De privacidad	18 %	16 %
Talkwalker	De precio	21 %	24 %
Talkwalker	En la comprensión del idioma	3 %	16 %
Talkwalker	Aporta información sesgada	0 %	12 %
Talkwalker	Aporta información engañosa	0 %	19 %
Talkwalker	Solo aporta información relativamente obvia	6 %	20 %
Talkwalker	Otras limitaciones	3 %	10 %
Digimind	En el desempeño o rendimiento	32 %	21 %
Digimind	En las funcionalidades	16 %	21 %
Digimind	En la compatibilidad con otros sistemas o con el equipo o dispositivo	16 %	14 %
Digimind	De seguridad	0 %	22 %
Digimind	De privacidad	26 %	13 %
Digimind	De precio	16 %	18 %
Digimind	En la comprensión del idioma	11 %	22 %
Digimind	Aporta información sesgada	37 %	21 %
Digimind	Aporta información engañosa	21 %	14 %
Digimind	Solo aporta información relativamente obvia	11 %	20 %
Digimind	Otras limitaciones	5 %	6 %
Resonate	En el desempeño o rendimiento	9 %	19 %
Resonate	En las funcionalidades	18 %	20 %
Resonate	En la compatibilidad con otros sistemas o con el equipo o dispositivo	9 %	20 %
Resonate	De seguridad	14 %	26 %
Resonate	De privacidad	9 %	21 %
Resonate	De precio	9 %	19 %
Resonate	En la comprensión del idioma	9 %	24 %
Resonate	Aporta información sesgada	27 %	15 %
Resonate	Aporta información engañosa	9 %	12 %
Resonate	Solo aporta información relativamente obvia	5 %	17 %
Resonate	Otras limitaciones	14 %	4 %
Sysomos	En el desempeño o rendimiento	0 %	15 %
Sysomos	En las funcionalidades	26 %	18 %
Sysomos	En la compatibilidad con otros sistemas o con el equipo o dispositivo	32 %	18 %
Sysomos	De seguridad	21 %	19 %
Sysomos	De privacidad	0 %	17 %
Sysomos	De precio	16 %	20 %
Sysomos	En la comprensión del idioma	16 %	23 %
Sysomos	Aporta información sesgada	16 %	18 %
Sysomos	Aporta información engañosa	5 %	19 %
Sysomos	Solo aporta información relativamente obvia	5 %	18 %
Sysomos	Otras limitaciones	11 %	10 %

Tabla 37: Limitaciones de asistentes virtuales.

Producto	**Tipo de limitaciones**	**Español**	**Inglés**
Google Assistant	En el desempeño o rendimiento	9 %	12 %
Google Assistant	En el desempeño o rendimiento	9 %	12 %
Google Assistant	En las funcionalidades	15 %	17 %
Google Assistant	En la compatibilidad con otros sistemas o con el equipo o dispositivo	7 %	9 %
Google Assistant	De seguridad	9 %	12 %
Google Assistant	De privacidad	15 %	16 %
Google Assistant	De precio	4 %	9 %
Google Assistant	En la comprensión del idioma	13 %	9 %
Google Assistant	Proporciona contenidos ofensivos, tóxicos o inadecuados	3 %	4 %
Google Assistant	Proporciona contenidos sesgados y/o estereotipados	7 %	5 %
Google Assistant	Proporciona información convincente a primera vista, pero errónea al verificarla	7 %	7 %
Google Assistant	Sólo ofrece información genérica que ya se conocía	17 %	9 %
Google Assistant	No justifica las respuestas ni da acceso a las fuentes que la justifican	12 %	9 %
Google Assistant	Otras limitaciones	6 %	3 %

Siri	En el desempeño o rendimiento	11 %	18 %
Siri	En las funcionalidades	15 %	14 %
Siri	En la compatibilidad con otros sistemas o con el equipo o dispositivo	13 %	12 %
Siri	De seguridad	10 %	10 %
Siri	De privacidad	18 %	11 %
Siri	De precio	10 %	7 %
Siri	En la comprensión del idioma	11 %	11 %
Siri	Proporciona contenidos ofensivos, tóxicos o inadecuados	2 %	5 %
Siri	Proporciona contenidos sesgados y/o estereotipados	7 %	6 %
Siri	Proporciona información convincente a primera vista, pero errónea al verificarla	4 %	11 %
Siri	Sólo ofrece información genérica que ya se conocía	19 %	13 %
Siri	No justifica las respuestas ni da acceso a las fuentes que la justifican	10 %	9 %
Siri	Otras limitaciones	4 %	4 %
Alexa	En el desempeño o rendimiento	11 %	16 %
Alexa	En las funcionalidades	14 %	16 %
Alexa	En la compatibilidad con otros sistemas o con el equipo o dispositivo	9 %	11 %
Alexa	De seguridad	9 %	11 %
Alexa	De privacidad	21 %	15 %
Alexa	De precio	9 %	9 %
Alexa	En la comprensión del idioma	11 %	12 %
Alexa	Proporciona contenidos ofensivos, tóxicos o inadecuados	2 %	5 %
Alexa	Proporciona contenidos sesgados y/o estereotipados	8 %	5 %
Alexa	Proporciona información convincente a primera vista, pero errónea al verificarla	8 %	8 %
Alexa	Sólo ofrece información genérica que ya se conocía	19 %	14 %
Alexa	No justifica las respuestas ni da acceso a las fuentes que la justifican	10 %	8 %
Alexa	Otras limitaciones	7 %	6 %
Bixby	En el desempeño o rendimiento	4 %	21 %
Bixby	En las funcionalidades	21 %	21 %
Bixby	En la compatibilidad con otros sistemas o con el equipo o dispositivo	12 %	13 %
Bixby	De seguridad	13 %	11 %
Bixby	De privacidad	9 %	12 %
Bixby	De precio	6 %	9 %
Bixby	En la comprensión del idioma	15 %	16 %
Bixby	Proporciona contenidos ofensivos, tóxicos o inadecuados	3 %	7 %
Bixby	Proporciona contenidos sesgados y/o estereotipados	6 %	6 %
Bixby	Proporciona información convincente a primera vista, pero errónea al verificarla	7 %	16 %
Bixby	Sólo ofrece información genérica que ya se conocía	15 %	16 %
Bixby	No justifica las respuestas ni da acceso a las fuentes que la justifican	10 %	11 %
Bixby	Otras limitaciones	10 %	6 %
Kore.ai	En el desempeño o rendimiento	19 %	17 %
Kore.ai	En las funcionalidades	19 %	12 %
Kore.ai	En la compatibilidad con otros sistemas o con el equipo o dispositivo	6 %	14 %
Kore.ai	De seguridad	31 %	19 %
Kore.ai	De privacidad	6 %	15 %
Kore.ai	De precio	19 %	15 %
Kore.ai	En la comprensión del idioma	12 %	15 %
Kore.ai	Proporciona contenidos ofensivos, tóxicos o inadecuados	12 %	14 %
Kore.ai	Proporciona contenidos sesgados y/o estereotipados	6 %	12 %
Kore.ai	Proporciona información convincente a primera vista, pero errónea al verificarla	19 %	12 %
Kore.ai	Sólo ofrece información genérica que ya se conocía	12 %	8 %
Kore.ai	No justifica las respuestas ni da acceso a las fuentes que la justifican	25 %	10 %
Kore.ai	Otras limitaciones	6 %	7 %
IBM Watson Assistant	En el desempeño o rendimiento	6 %	15 %
IBM Watson Assistant	En las funcionalidades	25 %	21 %
IBM Watson Assistant	En la compatibilidad con otros sistemas o con el equipo o dispositivo	6 %	21 %
IBM Watson Assistant	De seguridad	25 %	14 %
IBM Watson Assistant	De privacidad	6 %	11 %
IBM Watson Assistant	De precio	0 %	13 %
IBM Watson Assistant	En la comprensión del idioma	0 %	18 %
IBM Watson Assistant	Proporciona contenidos ofensivos, tóxicos o inadecuados	6 %	13 %
IBM Watson Assistant	Proporciona contenidos sesgados y/o estereotipados	19 %	10 %
IBM Watson Assistant	Proporciona información convincente a primera vista, pero errónea al verificarla	12 %	10 %
IBM Watson Assistant	Sólo ofrece información genérica que ya se conocía	12 %	8 %
IBM Watson Assistant	No justifica las respuestas ni da acceso a las fuentes que la justifican	12 %	13 %
IBM Watson Assistant	Otras limitaciones	6 %	5 %

Amazon Lex	En el desempeño o rendimiento	8 %	18 %
Amazon Lex	En las funcionalidades	18 %	16 %
Amazon Lex	En la compatibilidad con otros sistemas o con el equipo o dispositivo	13 %	14 %
Amazon Lex	De seguridad	10 %	15 %
Amazon Lex	De privacidad	5 %	15 %
Amazon Lex	De precio	8 %	12 %
Amazon Lex	En la comprensión del idioma	15 %	20 %
Amazon Lex	Proporciona contenidos ofensivos, tóxicos o inadecuados	10 %	13 %
Amazon Lex	Proporciona contenidos sesgados y/o estereotipados	13 %	16 %
Amazon Lex	Proporciona información convincente a primera vista, pero errónea al verificarla	5 %	16 %
Amazon Lex	Sólo ofrece información genérica que ya se conocía	10 %	8 %
Amazon Lex	No justifica las respuestas ni da acceso a las fuentes que la justifican	15 %	13 %
Amazon Lex	Otras limitaciones	5 %	5 %
Google Dialogflow	En el desempeño o rendimiento	28 %	14 %
Google Dialogflow	En las funcionalidades	22 %	21 %
Google Dialogflow	En la compatibilidad con otros sistemas o con el equipo o dispositivo	11 %	21 %
Google Dialogflow	De seguridad	14 %	16 %
Google Dialogflow	De privacidad	19 %	6 %
Google Dialogflow	De precio	8 %	22 %
Google Dialogflow	En la comprensión del idioma	11 %	15 %
Google Dialogflow	Proporciona contenidos ofensivos, tóxicos o inadecuados	14 %	11 %
Google Dialogflow	Proporciona contenidos sesgados y/o estereotipados	3 %	16 %
Google Dialogflow	Proporciona información convincente a primera vista, pero errónea al verificarla	19 %	13 %
Google Dialogflow	Sólo ofrece información genérica que ya se conocía	11 %	14 %
Google Dialogflow	No justifica las respuestas ni da acceso a las fuentes que la justifican	17 %	8 %
Google Dialogflow	Otras limitaciones	6 %	9 %
Amelia	En el desempeño o rendimiento	10 %	20 %
Amelia	En las funcionalidades	5 %	25 %
Amelia	En la compatibilidad con otros sistemas o con el equipo o dispositivo	10 %	15 %
Amelia	De seguridad	15 %	16 %
Amelia	De privacidad	15 %	8 %
Amelia	De precio	10 %	23 %
Amelia	En la comprensión del idioma	5 %	13 %
Amelia	Proporciona contenidos ofensivos, tóxicos o inadecuados	15 %	10 %
Amelia	Proporciona contenidos sesgados y/o estereotipados	5 %	16 %
Amelia	Proporciona información convincente a primera vista, pero errónea al verificarla	10 %	21 %
Amelia	Sólo ofrece información genérica que ya se conocía	10 %	16 %
Amelia	No justifica las respuestas ni da acceso a las fuentes que la justifican	10 %	11 %
Amelia	Otras limitaciones	10 %	10 %
ChatGPT	En el desempeño o rendimiento	9 %	16 %
ChatGPT	En las funcionalidades	11 %	14 %
ChatGPT	En la compatibilidad con otros sistemas o con el equipo o dispositivo	6 %	9 %
ChatGPT	De seguridad	9 %	12 %
ChatGPT	De privacidad	14 %	14 %
ChatGPT	De precio	12 %	12 %
ChatGPT	En la comprensión del idioma	9 %	8 %
ChatGPT	Proporciona contenidos ofensivos, tóxicos o inadecuados	5 %	9 %
ChatGPT	Proporciona contenidos sesgados y/o estereotipados	14 %	10 %
ChatGPT	Proporciona información convincente a primera vista, pero errónea al verificarla	26 %	18 %
ChatGPT	Sólo ofrece información genérica que ya se conocía	12 %	13 %
ChatGPT	No justifica las respuestas ni da acceso a las fuentes que la justifican	20 %	9 %
ChatGPT	Otras limitaciones	8 %	5 %
Google Bard	En el desempeño o rendimiento	20 %	13 %
Google Bard	En las funcionalidades	17 %	18 %
Google Bard	En la compatibilidad con otros sistemas o con el equipo o dispositivo	7 %	14 %
Google Bard	De seguridad	9 %	10 %
Google Bard	De privacidad	18 %	14 %
Google Bard	De precio	8 %	12 %
Google Bard	En la comprensión del idioma	16 %	11 %
Google Bard	Proporciona contenidos ofensivos, tóxicos o inadecuados	7 %	11 %
Google Bard	Proporciona contenidos sesgados y/o estereotipados	16 %	12 %
Google Bard	Proporciona información convincente a primera vista, pero errónea al verificarla	25 %	14 %
Google Bard	Sólo ofrece información genérica que ya se conocía	18 %	15 %
Google Bard	No justifica las respuestas ni da acceso a las fuentes que la justifican	16 %	10 %
Google Bard	Otras limitaciones	13 %	3 %

Tabla 38: Limitaciones de traducción automática.

		español	inglés
Google Translate	En el desempeño o rendimiento	8 %	13 %
Google Translate	En las funcionalidades	6 %	9 %
Google Translate	En la compatibilidad con otros sistemas o con el equipo o dispositivo	3 %	6 %
Google Translate	De seguridad	4 %	8 %
Google Translate	De privacidad	7 %	5 %
Google Translate	De precio	3 %	6 %
Google Translate	En un ámbito o sector específico (por ejemplo, traducción jurídica)	13 %	14 %
Google Translate	Presenta sesgos sistemáticos de traducción y/o estereotipados	16 %	9 %
Google Translate	Proporciona traducciones naturales y fluidas pero incorrectas	17 %	16 %
Google Translate	Traducciones literales que no captura el trasfondo del texto	45 %	21 %
Google Translate	Otras limitaciones	4 %	3 %
DeepL	En el desempeño o rendimiento	4 %	20 %
DeepL	En las funcionalidades	9 %	19 %
DeepL	En la compatibilidad con otros sistemas o con el equipo o dispositivo	4 %	6 %
DeepL	De seguridad	4 %	20 %
DeepL	De privacidad	4 %	15 %
DeepL	De precio	16 %	13 %
DeepL	En un ámbito o sector específico (por ejemplo, traducción jurídica)	12 %	15 %
DeepL	Presenta sesgos sistemáticos de traducción y/o estereotipados	9 %	15 %
DeepL	Proporciona traducciones naturales y fluidas pero incorrectas	12 %	10 %
DeepL	Traducciones literales que no captura el trasfondo del texto	12 %	19 %
DeepL	Otras limitaciones	11 %	11 %
Bing Translator o Microsoft Translator	En el desempeño o rendimiento	13 %	17 %
Bing Translator o Microsoft Translator	En las funcionalidades	14 %	18 %
Bing Translator o Microsoft Translator	En la compatibilidad con otros sistemas o con el equipo o dispositivo	7 %	13 %
Bing Translator o Microsoft Translator	De seguridad	8 %	13 %
Bing Translator o Microsoft Translator	De privacidad	8 %	12 %
Bing Translator o Microsoft Translator	De precio	8 %	10 %
Bing Translator o Microsoft Translator	En un ámbito o sector específico (por ejemplo, traducción jurídica)	17 %	16 %
Bing Translator o Microsoft Translator	Presenta sesgos sistemáticos de traducción y/o estereotipados	16 %	16 %
Bing Translator o Microsoft Translator	Proporciona traducciones naturales y fluidas pero incorrectas	14 %	16 %
Bing Translator o Microsoft Translator	Traducciones literales que no captura el trasfondo del texto	29 %	20 %
Bing Translator o Microsoft Translator	Otras limitaciones	8 %	3 %
Amazon Translate	En el desempeño o rendimiento	9 %	20 %
Amazon Translate	En las funcionalidades	15 %	13 %
Amazon Translate	En la compatibilidad con otros sistemas o con el equipo o dispositivo	6 %	12 %
Amazon Translate	De seguridad	9 %	15 %
Amazon Translate	De privacidad	9 %	16 %
Amazon Translate	De precio	9 %	12 %
Amazon Translate	En un ámbito o sector específico (por ejemplo, traducción jurídica)	15 %	16 %
Amazon Translate	Presenta sesgos sistemáticos de traducción y/o estereotipados	6 %	10 %
Amazon Translate	Proporciona traducciones naturales y fluidas pero incorrectas	15 %	12 %
Amazon Translate	Traducciones literales que no captura el trasfondo del texto	22 %	12 %
Amazon Translate	Otras limitaciones	2 %	3 %
Systran Translate	En el desempeño o rendimiento	13 %	19 %
Systran Translate	En las funcionalidades	6 %	16 %
Systran Translate	En la compatibilidad con otros sistemas o con el equipo o dispositivo	13 %	14 %
Systran Translate	De seguridad	23 %	12 %
Systran Translate	De privacidad	6 %	16 %
Systran Translate	De precio	26 %	16 %
Systran Translate	En un ámbito o sector específico (por ejemplo, traducción jurídica)	10 %	18 %
Systran Translate	Presenta sesgos sistemáticos de traducción y/o estereotipados	16 %	18 %
Systran Translate	Proporciona traducciones naturales y fluidas pero incorrectas	23 %	19 %
Systran Translate	Traducciones literales que no captura el trasfondo del texto	0 %	19 %
Systran Translate	Otras limitaciones	3 %	6 %
Reverso Translator	En el desempeño o rendimiento	11 %	16 %
Reverso Translator	En las funcionalidades	8 %	16 %
Reverso Translator	En la compatibilidad con otros sistemas o con el equipo o dispositivo	4 %	18 %
Reverso Translator	De seguridad	14 %	17 %
Reverso Translator	De privacidad	5 %	24 %
Reverso Translator	De precio	0 %	11 %

Reverso Translator	En un ámbito o sector específico (por ejemplo, traducción jurídica)	14 %	22 %
Reverso Translator	Presenta sesgos sistemáticos de traducción y/o estereotipados	19 %	11 %
Reverso Translator	Proporciona traducciones naturales y fluidas pero incorrectas	5 %	13 %
Reverso Translator	Traducciones literales que no captura el trasfondo del texto	33 %	16 %
Reverso Translator	Otras limitaciones	10 %	6 %
memoQ Translator PRO	En el desempeño o rendimiento	17 %	18 %
memoQ Translator PRO	En las funcionalidades	17 %	15 %
memoQ Translator PRO	En la compatibilidad con otros sistemas o con el equipo o dispositivo	8 %	16 %
memoQ Translator PRO	De seguridad	21 %	16 %
memoQ Translator PRO	De privacidad	4 %	22 %
memoQ Translator PRO	De precio	25 %	18 %
memoQ Translator PRO	En un ámbito o sector específico (por ejemplo, traducción jurídica)	12 %	11 %
memoQ Translator PRO	Presenta sesgos sistemáticos de traducción y/o estereotipados	17 %	11 %
memoQ Translator PRO	Proporciona traducciones naturales y fluidas pero incorrectas	21 %	12 %
memoQ Translator PRO	Traducciones literales que no captura el trasfondo del texto	8 %	19 %
memoQ Translator PRO	Otras limitaciones	4 %	3 %
Smartling	En el desempeño o rendimiento	8 %	19 %
Smartling	En las funcionalidades	27 %	21 %
Smartling	En la compatibilidad con otros sistemas o con el equipo o dispositivo	19 %	14 %
Smartling	De seguridad	12 %	20 %
Smartling	De privacidad	8 %	18 %
Smartling	De precio	15 %	11 %
Smartling	En un ámbito o sector específico (por ejemplo, traducción jurídica)	8 %	19 %
Smartling	Presenta sesgos sistemáticos de traducción y/o estereotipados	23 %	25 %
Smartling	Proporciona traducciones naturales y fluidas pero incorrectas	27 %	18 %
Smartling	Traducciones literales que no captura el trasfondo del texto	12 %	16 %
Smartling	Otras limitaciones	4 %	8 %
Crowdin	En el desempeño o rendimiento	0 %	24 %
Crowdin	En las funcionalidades	21 %	11 %
Crowdin	En la compatibilidad con otros sistemas o con el equipo o dispositivo	21 %	12 %
Crowdin	De seguridad	14 %	16 %
Crowdin	De privacidad	14 %	22 %
Crowdin	De precio	21 %	16 %
Crowdin	En un ámbito o sector específico (por ejemplo, traducción jurídica)	14 %	22 %
Crowdin	Presenta sesgos sistemáticos de traducción y/o estereotipados	21 %	15 %
Crowdin	Proporciona traducciones naturales y fluidas pero incorrectas	21 %	23 %
Crowdin	Traducciones literales que no captura el trasfondo del texto	14 %	18 %
Crowdin	Otras limitaciones	21 %	3 %
TextUnited	En el desempeño o rendimiento	5 %	20 %
TextUnited	En las funcionalidades	16 %	18 %
TextUnited	En la compatibilidad con otros sistemas o con el equipo o dispositivo	21 %	15 %
TextUnited	De seguridad	11 %	11 %
TextUnited	De privacidad	5 %	20 %
TextUnited	De precio	5 %	14 %
TextUnited	En un ámbito o sector específico (por ejemplo, traducción jurídica)	16 %	16 %
TextUnited	Presenta sesgos sistemáticos de traducción y/o estereotipados	16 %	19 %
TextUnited	Proporciona traducciones naturales y fluidas pero incorrectas	21 %	18 %
TextUnited	Traducciones literales que no captura el trasfondo del texto	16 %	12 %
TextUnited	Otras limitaciones	5 %	8 %
ChatGPT	En el desempeño o rendimiento	9 %	18 %
ChatGPT	En las funcionalidades	10 %	18 %
ChatGPT	En la compatibilidad con otros sistemas o con el equipo o dispositivo	7 %	12 %
ChatGPT	De seguridad	9 %	19 %
ChatGPT	De privacidad	13 %	14 %
ChatGPT	De precio	9 %	12 %
ChatGPT	En un ámbito o sector específico (por ejemplo, traducción jurídica)	10 %	13 %
ChatGPT	Presenta sesgos sistemáticos de traducción y/o estereotipados	14 %	15 %
ChatGPT	Proporciona traducciones naturales y fluidas pero incorrectas	18 %	15 %
ChatGPT	Traducciones literales que no captura el trasfondo del texto	16 %	16 %
ChatGPT	Otras limitaciones	7 %	8 %
Google Bard	En el desempeño o rendimiento	14 %	22 %
Google Bard	En las funcionalidades	8 %	18 %
Google Bard	En la compatibilidad con otros sistemas o con el equipo o dispositivo	8 %	18 %
Google Bard	De seguridad	15 %	17 %
Google Bard	De privacidad	14 %	16 %

Google Bard	De precio	9 %	10 %
Google Bard	En un ámbito o sector específico (por ejemplo, traducción jurídica)	11 %	18 %
Google Bard	Presenta sesgos sistemáticos de traducción y/o estereotipados	12 %	14 %
Google Bard	Proporciona traducciones naturales y fluidas pero incorrectas	24 %	19 %
Google Bard	Traducciones literales que no captura el trasfondo del texto	20 %	16 %
Google Bard	Otras limitaciones	9 %	6 %

Tabla 39: Limitaciones de teclados predictivos.

Producto	**Tipo de limitaciones**	**Español**	**Inglés**
Microsoft SwiftKey	En el desempeño o rendimiento	7 %	13 %
Microsoft SwiftKey	En las funcionalidades	9 %	14 %
Microsoft SwiftKey	En la compatibilidad con otros sistemas o con el equipo o dispositivo	5 %	14 %
Microsoft SwiftKey	De seguridad	6 %	14 %
Microsoft SwiftKey	De privacidad	9 %	14 %
Microsoft SwiftKey	De precio	1 %	10 %
Microsoft SwiftKey	Los términos que sugiere están sesgados y/o responden a estereotipos	11 %	20 %
Microsoft SwiftKey	Sugiere palabras que “suenan bien”, pero no son las más apropiadas	16 %	19 %
Microsoft SwiftKey	Lo que sugiere es previsible y convencional	19 %	19 %
Microsoft SwiftKey	Otras limitaciones	13 %	8 %
GBoard	En el desempeño o rendimiento	8 %	14 %
GBoard	En las funcionalidades	6 %	11 %
GBoard	En la compatibilidad con otros sistemas o con el equipo o dispositivo	4 %	15 %
GBoard	De seguridad	8 %	14 %
GBoard	De privacidad	8 %	17 %
GBoard	De precio	2 %	12 %
GBoard	Los términos que sugiere están sesgados y/o responden a estereotipos	7 %	11 %
GBoard	Sugiere palabras que “suenan bien”, pero no son las más apropiadas	15 %	27 %
GBoard	Lo que sugiere es previsible y convencional	20 %	14 %
GBoard	Otras limitaciones	6 %	7 %
Grammarly	En el desempeño o rendimiento	6 %	18 %
Grammarly	En las funcionalidades	12 %	11 %
Grammarly	En la compatibilidad con otros sistemas o con el equipo o dispositivo	17 %	15 %
Grammarly	De seguridad	10 %	11 %
Grammarly	De privacidad	10 %	11 %
Grammarly	De precio	21 %	17 %
Grammarly	Los términos que sugiere están sesgados y/o responden a estereotipos	10 %	9 %
Grammarly	Sugiere palabras que “suenan bien”, pero no son las más apropiadas	6 %	22 %
Grammarly	Lo que sugiere es previsible y convencional	10 %	15 %
Grammarly	Otras limitaciones	6 %	6 %
Fleksy	En el desempeño o rendimiento	12 %	17 %
Fleksy	En las funcionalidades	18 %	17 %
Fleksy	En la compatibilidad con otros sistemas o con el equipo o dispositivo	35 %	11 %
Fleksy	De seguridad	18 %	15 %
Fleksy	De privacidad	24 %	23 %
Fleksy	De precio	29 %	17 %
Fleksy	Los términos que sugiere están sesgados y/o responden a estereotipos	18 %	21 %
Fleksy	Sugiere palabras que “suenan bien”, pero no son las más apropiadas	29 %	25 %
Fleksy	Lo que sugiere es previsible y convencional	12 %	16 %
Fleksy	Otras limitaciones	6 %	4 %
iPhone	En el desempeño o rendimiento	10 %	14 %
iPhone	En las funcionalidades	8 %	11 %
iPhone	En la compatibilidad con otros sistemas o con el equipo o dispositivo	8 %	10 %
iPhone	De seguridad	5 %	9 %
iPhone	De privacidad	7 %	7 %
iPhone	De precio	7 %	9 %
iPhone	Los términos que sugiere están sesgados y/o responden a estereotipos	9 %	10 %
iPhone	Sugiere palabras que “suenan bien”, pero no son las más apropiadas	19 %	22 %
iPhone	Lo que sugiere es previsible y convencional	17 %	18 %
iPhone	Otras limitaciones	9 %	6 %
GMail	En el desempeño o rendimiento	7 %	14 %
GMail	En las funcionalidades	6 %	10 %
GMail	En la compatibilidad con otros sistemas o con el equipo o dispositivo	5 %	9 %

GMail	De seguridad	5 %	11 %
GMail	De privacidad	9 %	11 %
GMail	De precio	2 %	9 %
GMail	Los términos que sugiere están sesgados y/o responden a estereotipos	6 %	6 %
GMail	Sugiere palabras que “suenan bien”, pero no son las más apropiadas	13 %	15 %
GMail	Lo que sugiere es previsible y convencional	22 %	18 %
GMail	Otras limitaciones	4 %	7 %
Google Workspaces	En el desempeño o rendimiento	6 %	15 %
Google Workspaces	En las funcionalidades	12 %	15 %
Google Workspaces	En la compatibilidad con otros sistemas o con el equipo o dispositivo	7 %	13 %
Google Workspaces	De seguridad	9 %	13 %
Google Workspaces	De privacidad	16 %	16 %
Google Workspaces	De precio	12 %	12 %
Google Workspaces	Los términos que sugiere están sesgados y/o responden a estereotipos	12 %	9 %
Google Workspaces	Sugiere palabras que “suenan bien”, pero no son las más apropiadas	17 %	15 %
Google Workspaces	Lo que sugiere es previsible y convencional	13 %	16 %
Google Workspaces	Otras limitaciones	6 %	4 %
Microsoft Outlook	En el desempeño o rendimiento	7 %	12 %
Microsoft Outlook	En las funcionalidades	8 %	11 %
Microsoft Outlook	En la compatibilidad con otros sistemas o con el equipo o dispositivo	6 %	13 %
Microsoft Outlook	De seguridad	5 %	10 %
Microsoft Outlook	De privacidad	7 %	10 %
Microsoft Outlook	De precio	7 %	9 %
Microsoft Outlook	Los términos que sugiere están sesgados y/o responden a estereotipos	8 %	10 %
Microsoft Outlook	Sugiere palabras que “suenan bien”, pero no son las más apropiadas	15 %	14 %
Microsoft Outlook	Lo que sugiere es previsible y convencional	23 %	17 %
Microsoft Outlook	Otras limitaciones	5 %	7 %
Microsoft Office 365	En el desempeño o rendimiento	6 %	15 %
Microsoft Office 365	En las funcionalidades	6 %	12 %
Microsoft Office 365	En la compatibilidad con otros sistemas o con el equipo o dispositivo	6 %	13 %
Microsoft Office 365	De seguridad	6 %	14 %
Microsoft Office 365	De privacidad	10 %	13 %
Microsoft Office 365	De precio	12 %	13 %
Microsoft Office 365	Los términos que sugiere están sesgados y/o responden a estereotipos	9 %	10 %
Microsoft Office 365	Sugiere palabras que “suenan bien”, pero no son las más apropiadas	10 %	15 %
Microsoft Office 365	Lo que sugiere es previsible y convencional	16 %	14 %
Microsoft Office 365	Otras limitaciones	7 %	5 %

Tabla 40: Limitaciones de buscadores web

Producto	**Tipo de limitaciones**	**Español**	**Inglés**
Google Search	En el desempeño o rendimiento	4 %	9 %
Google Search	En las funcionalidades	5 %	8 %
Google Search	En la compatibilidad con otros sistemas o con el equipo o dispositivo	3 %	6 %
Google Search	De seguridad	12 %	10 %
Google Search	De privacidad	18 %	13 %
Google Search	De precio	3 %	5 %
Google Search	Devuelve páginas con contenidos tóxicos, agresivos o inapropiados en general	4 %	6 %
Google Search	Sesgos / preferencias sistemáticas hacia determinados tipos de páginas	19 %	11 %
Google Search	Devuelve páginas relacionadas con la consulta pero que contienen información engañosa	12 %	9 %
Google Search	Devuelve páginas con información previsible con poco valor añadido	12 %	11 %
Google Search	Otras limitaciones	4 %	3 %
Bing	En el desempeño o rendimiento	13 %	12 %
Bing	En las funcionalidades	12 %	14 %
Bing	En la compatibilidad con otros sistemas o con el equipo o dispositivo	8 %	10 %
Bing	De seguridad	11 %	9 %
Bing	De privacidad	15 %	9 %
Bing	De precio	4 %	6 %
Bing	Devuelve páginas con contenidos tóxicos, agresivos o inapropiados en general	7 %	5 %
Bing	Sesgos / preferencias sistemáticas hacia determinados tipos de páginas	23 %	10 %
Bing	Devuelve páginas relacionadas con la consulta pero que contienen información engañosa	12 %	10 %
Bing	Devuelve páginas con información previsible con poco valor añadido	20 %	14 %
Bing	Otras limitaciones	6 %	2 %
Yahoo Search	En el desempeño o rendimiento	16 %	15 %
Yahoo Search	En las funcionalidades	16 %	13 %
Yahoo Search	En la compatibilidad con otros sistemas o con el equipo o dispositivo	10 %	9 %
Yahoo Search	De seguridad	14 %	9 %
Yahoo Search	De privacidad	14 %	9 %

Yahoo Search	De precio	3 %	5 %
Yahoo Search	Devuelve páginas con contenidos tóxicos, agresivos o inapropiados en general	8 %	6 %
Yahoo Search	Sesgos / preferencias sistemáticas hacia determinados tipos de páginas	20 %	11 %
Yahoo Search	Devuelve páginas relacionadas con la consulta pero que contienen información engañosa	17 %	10 %
Yahoo Search	Devuelve páginas con información previsible con poco valor añadido	21 %	13 %
Yahoo Search	Otras limitaciones	4 %	4 %
DuckDuckGo	En el desempeño o rendimiento	13 %	15 %
DuckDuckGo	En las funcionalidades	14 %	16 %
DuckDuckGo	En la compatibilidad con otros sistemas o con el equipo o dispositivo	17 %	13 %
DuckDuckGo	De seguridad	9 %	11 %
DuckDuckGo	De privacidad	6 %	12 %
DuckDuckGo	De precio	4 %	7 %
DuckDuckGo	Devuelve páginas con contenidos tóxicos, agresivos o inapropiados en general	7 %	7 %
DuckDuckGo	Sesgos / preferencias sistemáticas hacia determinados tipos de páginas	12 %	11 %
DuckDuckGo	Devuelve páginas relacionadas con la consulta pero que contienen información engañosa	14 %	10 %
DuckDuckGo	Devuelve páginas con información previsible con poco valor añadido	19 %	15 %
DuckDuckGo	Otras limitaciones	11 %	5 %
Brave Search	En el desempeño o rendimiento	19 %	22 %
Brave Search	En las funcionalidades	8 %	18 %
Brave Search	En la compatibilidad con otros sistemas o con el equipo o dispositivo	12 %	13 %
Brave Search	De seguridad	10 %	13 %
Brave Search	De privacidad	6 %	16 %
Brave Search	De precio	12 %	12 %
Brave Search	Devuelve páginas con contenidos tóxicos, agresivos o inapropiados en general	6 %	18 %
Brave Search	Sesgos / preferencias sistemáticas hacia determinados tipos de páginas	4 %	18 %
Brave Search	Devuelve páginas relacionadas con la consulta pero que contienen información engañosa	13 %	19 %
Brave Search	Devuelve páginas con información previsible con poco valor añadido	17 %	12 %
Brave Search	Otras limitaciones	8 %	6 %
Elasticsearch	En el desempeño o rendimiento	14 %	22 %
Elasticsearch	En las funcionalidades	0 %	22 %
Elasticsearch	En la compatibilidad con otros sistemas o con el equipo o dispositivo	14 %	19 %
Elasticsearch	De seguridad	7 %	19 %
Elasticsearch	De privacidad	29 %	13 %
Elasticsearch	De precio	29 %	16 %
Elasticsearch	Devuelve páginas con contenidos tóxicos, agresivos o inapropiados en general	7 %	22 %
Elasticsearch	Sesgos / preferencias sistemáticas hacia determinados tipos de páginas	7 %	16 %
Elasticsearch	Devuelve páginas relacionadas con la consulta pero que contienen información engañosa	0 %	18 %
Elasticsearch	Devuelve páginas con información previsible con poco valor añadido	21 %	12 %
Elasticsearch	Otras limitaciones	14 %	4 %
Mindbreeze	En el desempeño o rendimiento	27 %	26 %
Mindbreeze	En las funcionalidades	5 %	21 %
Mindbreeze	En la compatibilidad con otros sistemas o con el equipo o dispositivo	14 %	19 %
Mindbreeze	De seguridad	14 %	16 %
Mindbreeze	De privacidad	9 %	20 %
Mindbreeze	De precio	14 %	20 %
Mindbreeze	Devuelve páginas con contenidos tóxicos, agresivos o inapropiados en general	23 %	21 %
Mindbreeze	Sesgos / preferencias sistemáticas hacia determinados tipos de páginas	14 %	16 %
Mindbreeze	Devuelve páginas relacionadas con la consulta pero que contienen información engañosa	9 %	16 %
Mindbreeze	Devuelve páginas con información previsible con poco valor añadido	23 %	16 %
Mindbreeze	Otras limitaciones	9 %	3 %
Apache Solr	En el desempeño o rendimiento	10 %	20 %
Apache Solr	En las funcionalidades	10 %	17 %
Apache Solr	En la compatibilidad con otros sistemas o con el equipo o dispositivo	10 %	17 %
Apache Solr	De seguridad	25 %	18 %
Apache Solr	De privacidad	15 %	21 %
Apache Solr	De precio	30 %	16 %
Apache Solr	Devuelve páginas con contenidos tóxicos, agresivos o inapropiados en general	25 %	14 %
Apache Solr	Sesgos / preferencias sistemáticas hacia determinados tipos de páginas	20 %	13 %
Apache Solr	Devuelve páginas relacionadas con la consulta pero que contienen información engañosa	15 %	16 %
Apache Solr	Devuelve páginas con información previsible con poco valor añadido	15 %	14 %
Apache Solr	Otras limitaciones	0 %	5 %
Perplexity	En el desempeño o rendimiento	15 %	15 %
Perplexity	En las funcionalidades	7 %	20 %
Perplexity	En la compatibilidad con otros sistemas o con el equipo o dispositivo	15 %	24 %
Perplexity	De seguridad	7 %	23 %
Perplexity	De privacidad	33 %	18 %
Perplexity	De precio	7 %	20 %
Perplexity	Devuelve páginas con contenidos tóxicos, agresivos o inapropiados en general	7 %	23 %
Perplexity	Sesgos / preferencias sistemáticas hacia determinados tipos de páginas	15 %	27 %
Perplexity	Devuelve páginas relacionadas con la consulta pero que contienen información engañosa	11 %	12 %
Perplexity	Devuelve páginas con información previsible con poco valor añadido	19 %	18 %
Perplexity	Otras limitaciones	15 %	3 %

Tabla 32: Encuestas adopción.

Análisis de opiniones	español			inglés		
	personal	profesional	ambos	personal	profesional	ambos
Sprinklr	9	5	9	42	39	27
Khoros	16	10	5	46	28	26
NetBase Quid	9	4	2	44	29	20
Brandwatch	8	6	8	60	27	28
Linkfluence	12	10	4	54	31	40
Synthesio	10	3	4	47	35	21
Talkwalker	20	7	8	58	34	25
Digimind	9	6	7	43	37	22
Resonate	10	6	5	46	26	32
Meltwater	12	4	4	39	36	24
Asistentes virtuales	español			inglés		
	personal	profesional	ambos	personal	profesional	ambos
Google Assistant	271	16	43	231	30	74
Siri	206	10	32	234	30	70
Alexa	268	11	25	260	31	53
Bixby	50	10	6	104	34	25
Cortana	151	14	27	136	40	33
Kore.ai	4	5	4	36	22	20
IBM Watson Assistant	5	4	5	37	38	21
Amazon Lex	18	7	7	50	30	27
Google Dialogflow	18	10	10	40	38	32
Amelia	7	5	1	36	32	20
ChatGPT	41	16	18	57	38	25
Traducción automática	español			inglés		
	personal	profesional	ambos	personal	profesional	ambos
Google Translate	369	42	289	257	48	124
DeepL	24	23	59	30	33	30
Bing Translator o Microsoft Translator	56	17	23	82	47	30
Amazon Translate	38	8	15	74	38	36
Systran Translate	10	7	5	41	29	25
Reverso Translator	48	16	28	33	35	30
memoQ Translator PRO	3	5	9	33	35	20
Smartling	7	5	9	39	31	21
Crowdin	3	6	13	33	33	22
TextUnited	11	7	9	39	31	30
Teclados predictivos	español			inglés		
	personal	profesional	ambos	personal	profesional	ambos
Microsoft SwiftKey	79	8	24	66	37	49
GBoard	120	7	36	85	41	40
Grammarly Keyboard	16	18	12	121	49	90
Fleksy	6	5	4	31	26	21
iPhone Keyboard	127	14	67	204	45	112
GMail	183	26	123	193	49	131
Google Workspace	33	19	26	82	54	56
Microsoft Outlook	105	54	88	121	112	85
Microsoft Office 365	79	50	99	104	91	119
Buscadores web	español			inglés		
	personal	profesional	ambos	personal	profesional	ambos
Google Search	313	19	486	434	62	323
Bing	193	31	80	346	62	109
Yahoo Search	197	13	56	343	46	120
DuckDuckGo	61	2	21	180	41	38
Brave Search	26	6	13	66	18	36
Elasticsearch	5	4	6	35	26	28
Mindbreeze	2	3	4	35	22	24
Apache Solr	5	5	5	27	36	22

Tabla 33: Resultados RRSS.

Análisis de opiniones	español			inglés		
	positivo	neutral	negativo	positivo	neutral	negativo
Brandwatch	285	166	10	2264	2502	90
Digimind	131	35	3	141	254	30
Meltwater	105	46	7	1630	1033	124
NetBase Quid	185	80	0	5942	8900	147
Sprinklr	1625	2553	136	4519	4001	684
Talkwalker	101	26	6	2898	2098	222
Asistentes virtuales	español			inglés		
	positivo	neutral	negativo	positivo	neutral	negativo
Alexa	7694	2696	2122	7457	8089	9451
Bixby	1614	1347	342	3756	9685	1544
ChatGPT	7643	2185	2734	11542	7760	5848
Google Assistant	7043	2338	2822	8392	7593	4000
Google Bard	2367	711	1023	4080	6155	4799
Siri	3353	2871	1454	4367	8026	2605
Traducción automática	español			inglés		
	positivo	neutral	negativo	positivo	neutral	negativo
Microsoft Translator o Bing Translator	3698	778	489	6503	2766	3216
DeepL	3411	2115	1350	4285	11097	2554
Google Translate	4953	4504	3838	5843	9525	4687
memoQ Translator PRO	54	128	97	417	1455	136
Reverso Translator	1220	90	56	1516	232	201
Smartling	4	783	4	465	497	37
Teclados predictivos	español			inglés		
	positivo	neutral	negativo	positivo	neutral	negativo
Fleksy	187	43	82	756	580	544
GBoard	5472	2610	639	8361	7859	3875
GMail	3209	517	1670	3442	728	3227
Grammarly	151	19	89	6410	1498	2228
iPhone Keyboard	67	289	78	814	1122	255
Microsoft Office 365	5657	3031	1125	6910	11904	1186
Microsoft Outlook	4160	586	259	3465	720	960
Microsoft SwiftKey	4076	612	457	4288	1643	1142
Buscadores web	español			inglés		
	positivo	neutral	negativo	positivo	neutral	negativo
Bing	5104	898	1118	9366	6955	5364
Brave Search	4413	497	355	7238	2542	2071
DuckDuckGo	5133	2053	1255	7185	9531	4675
Elasticsearch	144	275	81	3332	10107	1620
Google Search	4973	2043	2239	10266	8335	1399
Perplexity	2071	3755	235	4100	11975	1178
Yahoo Search	66	69	82	2049	3043	1437

Tabla 34: Encuestas de satisfacción.

Análisis de opiniones	español	inglés
Sprinklr	3,73	4,01
Khoros	3,93	4,15
NetBase Quid	3,95	4,1
Brandwatch	3,82	4,32
Linkfluence	3,79	4,28
Synthesio	3,48	4,16
Talkwalker	3,61	4,22
Digimind	3,95	4,24
Resonate	3,64	4,05
Sysomos	4,0	4,23
Asistentes virtuales	español	inglés
Google Assistant	3,53	4,1
Siri	3,54	4,04
Alexa	3,71	4,1
Bixby	3,04	3,73
Kore.ai	3,5	4,2
IBM Watson Assistant	3,31	4,14
Amazon Lex	3,54	4,26
Google Dialogflow	3,53	4,21
186,Amelia	3,65	4,26
ChatGPT	3,85	4,19
Google Bard	3,43	4,1
Traducción automática	español	inglés
Google Translate	3,76	4,25
DeepL	4,19	4,23
Bing Translator o Microsoft Translator	3,38	3,98
Amazon Translate	3,69	4,21
Systran Translate	3,65	4,1
Reverso Translator	3,55	3,99
memoQ Translator PRO	3,46	4,19
Smartling	3,65	4,0
Crowdin	3,36	4,04
TextUnited	3,37	4,01
ChatGPT	3,85	4,23
Google Bard	3,53	4,18
Teclados predictivos	español	inglés
Microsoft SwiftKey	3,79	4,17
GBoard	3,74	4,15
Grammarly	3,83	4,17
Fleksy	3,59	4,08
iPhone	3,63	4,05
GMail	3,82	4,17
Google Workspaces	3,71	4,22
Microsoft Outlook	3,64	4,06
Microsoft Office 365	3,76	4,12
Buscadores web	español	inglés
Google Search	4,26	4,43
Bing	3,05	3,78
Yahoo Search	2,91	3,81
DuckDuckGo	3,62	4,12
Brave Search	3,73	3,81
Elasticsearch	3,21	4,06
Mindbreeze	3,59	4,11
Apache Solr	3,65	4,08
Perplexity	3,3	4,04

Tabla 35: Encuestas de limitaciones.

Análisis de opiniones	español	inglés
Sprinklr	0,13	0,17
Sysomos	0,13	0,18
Khoros	0,12	0,15
NetBase Quid	0,14	0,15
Brandwatch	0,1	0,17
Linkfluence	0,11	0,16
Synthesio	0,13	0,16
Talkwalker	0,07	0,17
Digimind	0,17	0,17
Resonate	0,12	0,18
Asistentes virtuales	español	inglés
Google Assistant	0,09	0,09
186,Amelia	0,1	0,16
ChatGPT	0,12	0,11
Google Bard	0,15	0,12
Siri	0,1	0,1
Alexa	0,11	0,1
Bixby	0,1	0,13
Kore.ai	0,15	0,13
IBM Watson Assistant	0,11	0,13
Amazon Lex	0,1	0,14
Google Dialogflow	0,14	0,14
Traducción automática	español	inglés
Google Translate	0,12	0,1
TextUnited	0,12	0,15
ChatGPT	0,11	0,14
Google Bard	0,13	0,16
DeepL	0,09	0,15
Bing Translator o Microsoft Translator	0,13	0,14
Amazon Translate	0,11	0,13
Systran Translate	0,13	0,16
Reverso Translator	0,11	0,15
memoQ Translator PRO	0,14	0,15
Smartling	0,15	0,17
Crowdin	0,17	0,16
Teclados predictivos	español	inglés
Microsoft SwiftKey	0,1	0,15
Microsoft Office 365	0,09	0,12
GBoard	0,08	0,14
Grammarly	0,11	0,13
Fleksy	0,2	0,17
iPhone	0,1	0,12
GMail	0,08	0,11
Google Workspaces	0,11	0,13
Microsoft Outlook	0,09	0,11
Buscadores web	español	inglés
Google Search	0,09	0,08
Bing	0,12	0,09
Yahoo Search	0,13	0,09
DuckDuckGo	0,11	0,11
Brave Search	0,1	0,15
Elasticsearch	0,13	0,17
Mindbreeze	0,14	0,18
Apache Solr	0,16	0,16
Perplexity	0,14	0,18