UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA (UNED)
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

MÉTODOS Y HERRAMIENTAS PARA LA EVALUACIÓN DE RESÚMENES AUTOMÁTICOS MEDIANTE FEEDBACK HUMANO

TRABAJO DE FIN DE MÁSTER PRESENTADO POR
DAVID CELESTINO COLÁS ROMANOS

DIRIGIDO POR
JUAN MANUEL CIGARRAN RECUERO
ENRIQUE AMIGÓ CABRERA

MÁSTER UNIVERSITARIO EN TECNOLOGÍAS DEL LENGUAJE
CURSO 2022-2023
CONVOCATORIA DE SEPTIEMBRE

ÍNDICE

MÉTODOS Y HERRAMIENTAS PARA LA EVALUACIÓN DE RESÚMENES AUTOMÁTICOS MEDIANTE FEEDBACK HUMANO

David Celestino Colás Romanos
Dpto. de Lenguajes y Sistemas Informáticos
UNED
davidcolroma@gmail.com
dcolas4@alumno.uned.es

Juan Manuel Cigarran Recuero
Dpto. de Lenguajes y Sistemas Informáticos
UNED
juanci@lsi.uned.es

Enrique Amigó Cabrera
Dpto. de Lenguajes y Sistemas Informáticos
UNED
enrique@lsi.uned.es

ABSTRACT

A medida que avanzan los modelos de lenguaje natural, el entrenamiento y la evaluación de estos se ven limitados por las métricas y los datos empleados para tareas específicas. En el contexto de la generación automática de resúmenes, comúnmente se utilizan métricas tradicionales como ROUGE y BLEU, entre otras, pero estas podrían no capturar la verdadera esencia de la calidad del resumen. En este trabajo, se revisa la validez de dichas métricas en el contexto actual, empleando un dataset innovador de OpenAI, compuesto por comparaciones de resúmenes anotados con feedback humano. Se observa que los resúmenes automáticos a menudo superan la calidad de los resúmenes de referencia humanos, llegando a ser casi indistinguibles de estos. Mediante diversos experimentos, se explora tanto la eficacia de las métricas de evaluación tradicionales como el impacto de ciertos rasgos y características en la calidad percibida de un resumen. Este estudio ofrece tres contribuciones significativas: en primer lugar, proporciona una evaluación crítica de las métricas estándar en el contexto actual, subrayando la necesidad de adaptaciones continuas. En segundo lugar, pone de relieve la importancia del feedback humano y cómo este puede enriquecer el proceso de evaluación, brindando percepciones valiosas que las métricas tradicionales podrían no capturar. Finalmente, introduce y valida métricas y herramientas innovadoras, como aquellas basadas en la similitud semántica, y la herramienta `ExplainSumm`, que han demostrado su eficacia en distintos contextos. Se espera que este trabajo no solo desafíe las nociones convencionales en la evaluación de resúmenes automáticos, sino que también proponga una ruta hacia una evaluación más holística y matizada, capaz de representar de manera más fiable la calidad y utilidad de los resúmenes en la era contemporánea.

***Keywords* —** Machine Learning (cs.LG), Computation and Language (cs.CL), Machine Learning (stat.ML), FOS: Computer and information sciences, FOS: Computer and information sciences

1. INTRODUCCIÓN

La generación automática de resúmenes ha experimentado avances significativos en los últimos años, impulsada principalmente por la emergencia de modelos de lenguaje generativos como GPT (Radford et al. 2018). Estos modelos, potenciados por técnicas supervisadas de entrenamiento y retroalimentación humana (human feedback), han establecido nuevos estándares de desempeño en la generación de resúmenes automáticos, acercándose e incluso superando en algunos casos, la calidad de resúmenes generados por humanos (Stiennon et al. 2020).

A pesar de estos avances, surgen nuevos desafíos en la evaluación de estos modelos. Históricamente, métricas como ROUGE y BLEU, entre otras, han sido pilares en la evaluación de resúmenes automáticos (Lin 2004; Papineni et al. 2002). Sin embargo, a medida que los resúmenes generados por modelos avanzan, estas métricas tradicionales muestran sus limitaciones, teniendo cada vez menos correlación con el juicio humano (Novikova et al. 2017; Schluter 2017). Además, la línea que distingue los resúmenes automáticos de los generados por humanos se vuelve cada vez más difusa, planteando cuestiones sobre la relevancia de los resúmenes de referencia en la evaluación (Stiennon et al. 2020).

Por tanto, surge una imperativa necesidad de explorar y adoptar nuevas métricas y metodologías de evaluación. Las métricas basadas en embeddings y la similitud semántica, por ejemplo, ofrecen prometedores enfoques que pueden estar más alineados con el juicio humano y abordar la complejidad inherente a la evaluación de resúmenes automáticos de calidad.

En este trabajo, nos embarcamos en una profunda investigación con el objetivo de comprender mejor el panorama actual de la evaluación de modelos de resumen automático, cuestionar las métricas tradicionales y proponer enfoques novedosos. Para ello, hemos empleado un dataset innovador de OpenAI, compuesto por comparaciones de resúmenes anotados con feedback humano, el cual se explicará en detalle en la Sección 3.2. Asimismo, buscamos entender cómo ciertos rasgos y características pueden influir en la calidad percibida de un resumen, y cómo herramientas como `ExplainSumm` pueden ofrecer percepciones valiosas en este dominio.

Esta investigación se estructura en torno a tres experimentos principales, cuyos resultados y discusiones se presentan en las secciones posteriores. A continuación, proporcionamos una breve descripción de la organización del trabajo:

En la Sección 2, abordamos el "Estado del Arte", revisando las investigaciones y avances que definen la dirección actual en el campo de la generación automática de resúmenes. La Sección 3 se adentra en la "Metodología y diseño experimental", ofreciendo detalles sobre cómo se planteó y desarrolló nuestra investigación, desde las bases metodológicas hasta la construcción y análisis de cada experimento. En la Sección 4, desplegamos los "Resultados y discusión", donde nuestros hallazgos son expuestos y contextualizados en relación con el dominio del resumen automático. Finalmente, la Sección 5 se centra en el "Trabajo futuro y Conclusiones", reflexionando sobre las aportaciones del estudio y planteando perspectivas para futuras investigaciones en el ámbito del resumen automático y su evaluación. Con ello, esperamos ofrecer una visión amplia y detallada del estado actual y los retos futuros de la generación y evaluación de resúmenes automáticos.

1.1. Objetivo e hipótesis

Frente a este escenario, surge una pregunta de investigación crucial: **¿qué métricas o rasgos, dado el contexto actual, pueden identificar efectivamente la calidad de los resúmenes?** Para abordar esta cuestión, pretendemos entender la importancia de contextualizar las métricas en el tiempo, y determinar qué métricas y rasgos son explicativos de la decisión del anotador a la hora de decidir cuando un resumen es mejor que otro.

1.2. Preguntas de investigación

Abordaremos el problema en varias etapas, cada una de las cuales abordará las siguientes preguntas:

- **PI1**: ¿Son aquellos resúmenes generados por modelos más complejos (arquitectura / número de parámetros), propios del contexto actual, indistinguibles de resúmenes de referencia o incluso mejores?
- **PI2**: ¿Son válidas las métricas de evaluación tradicionales (ROUGE, BLEU) en el contexto actual, en el cual el evaluador no es capaz de distinguir resúmenes automáticos de resúmenes de referencia (juicio humano)?
- **PI3**: ¿Es la propia calidad de los resúmenes a comparar una causa directa de la pérdida de capacidad predictiva de ciertos rasgos o métricas?
- **PI4**: ¿Qué métricas o rasgos son capaces de mantener su capacidad predictiva conforme se escala la calidad de los resúmenes?

1.3. Contexto de la Investigación

La tarea de evaluar la calidad de los resúmenes automáticos ha sido un área de estudio activa en la comunidad de procesamiento del lenguaje natural. A medida que las tecnologías y algoritmos avanzan, las métricas tradicionales utilizadas para medir la calidad de los resúmenes, basadas en comparaciones superficiales con referencias humanas, pueden no reflejar adecuadamente la calidad y utilidad real de los resúmenes generados. Además, al depender en gran medida de referencias humanas, se corre el riesgo de no capturar la verdadera esencia y riqueza de los resúmenes producidos por sistemas automatizados que pueden ser distintos, pero igualmente válidos.

En este contexto, identificamos dos limitaciones principales en la literatura existente. Primero, muchos de los estudios previos (Chaganty et al. 2018; Novikova et al. 2017) han enfocado sus métricas de evaluación en un contexto local, generalmente basando la calidad del resumen en su correlación con métricas como la adecuación, la cobertura, la coherencia y la precisión en comparación con un resumen de referencia. Si bien estas métricas pueden proporcionar una visión válida sobre la calidad de un resumen, también tienen el potencial de ser miopes y no capturar la complejidad y diversidad de resúmenes de alta calidad generados por modelos más avanzados.

Segundo, ha habido un supuesto implícito en la literatura de que la eficacia de los algoritmos de resúmenes automáticos es estática y no evoluciona con el tiempo (Fabbri et al. 2020). Sin embargo, en la era de los modelos de aprendizaje profundo y el constante avance de las técnicas de generación automática, este supuesto ya no es válido.

Con estas limitaciones en mente, este trabajo busca abordar la evaluación de resúmenes automáticos desde dos perspectivas esenciales y complementarias:

- **Evolución temporal de la eficacia de los algoritmos**: A diferencia de la suposición de que la calidad de los algoritmos de generación de resúmenes es estática, en esta investigación reconocemos que la eficacia de estos algoritmos cambia y evoluciona con el tiempo. Factores como la complejidad creciente de los algoritmos, el aumento en el número de parámetros y las innovaciones en las arquitecturas y técnicas de entrenamiento pueden influir en la calidad y naturaleza de los resúmenes generados.
- **Alineamiento humano**: Dado que los humanos son los principales consumidores de resúmenes, es esencial considerar cómo los resúmenes generados por máquinas se alinean con las evaluaciones humanas. Esto no solo se refiere a la comparación directa con referencias humanas, sino también a cómo los humanos perciben y valoran los resúmenes generados automáticamente en términos de utilidad, coherencia, informatividad y legibilidad.

A través de estas perspectivas, buscamos proporcionar una visión más holística y contemporánea de la evaluación de resúmenes automáticos, alejándonos de las métricas tradicionales y buscando una evaluación que sea más representativa de la utilidad y calidad real de los resúmenes en el mundo real.

1.4. Contribuciones

Este trabajo trasciende la mera investigación teórica para presentar contribuciones tangibles que buscan reformular el entendimiento y enfoque de la generación y evaluación de resúmenes automáticos. A continuación, se destacan las innovaciones y hallazgos más significativos derivados de nuestro estudio:

(1) **Reevaluación de las métricas de evaluación tradicionales:** Este estudio confirma que las métricas tradicionales, como ROUGE y BLEU, no son efectivas en el panorama actual de la generación automática de resúmenes. Nuestra investigación respalda y enriquece la discusión actual, destacando la imperativa necesidad de reevaluar y adaptar continuamente nuestras aproximaciones. Además, nuestro trabajo indica que, en muchos casos, los resúmenes generados por modelos avanzados son indistinguibles o incluso preferidos sobre los resúmenes humanos.

(2) **Integración y valoración del feedback humano:** Hemos reiterado la importancia del feedback humano en la evaluación de resúmenes automáticos. Nuestros hallazgos en el Experimento 2 (ver Sección 3.3.2) están en sintonía con la creciente atención que la comunidad ha dado al feedback humano. En particular, nuestros resultados sugieren una preferencia por conjuntos de

datos anotados con un enfoque comparativo en lugar de anotaciones tipo Likert[1], lo que resalta la riqueza y variabilidad que estos pueden aportar.

(3) **Introducción y validación de nuevas métricas y herramientas:** Hemos identificado la métrica de similitud del coseno entre los embeddings del texto y el resumen usando la librería `SentenceTransformer`, como una prometedora dirección de investigación para la evaluación de resúmenes automáticos. Además, `ExplainSumm` (ver Sección 3.2.4), el método propuesto, ha demostrado ser eficiente al identificar las métricas y rasgos más relevantes, siendo resiliente en diferentes escenarios de calidad de resúmenes.

(4) **Insights sobre rasgos intrínsecos y métricas:** El Experimento 3 (ver Sección 3.3.3) ha resaltado que, además de las métricas de evaluación, los rasgos inherentes de los resúmenes son fundamentales para la evaluación de la calidad. Observaciones como la influencia del número de palabras o monosílabos en el resumen y su relación inversa con la calidad subrayan la complejidad de la evaluación y sugieren una aproximación más holística y matizada en investigaciones futuras.

En resumen, este estudio no solo desafía y replantea las métricas convencionales, sino que también pone de manifiesto el valor del juicio humano y propone nuevas herramientas para un análisis más profundo y holístico de la generación automática de resúmenes. Las contribuciones aquí presentadas sientan las bases para investigaciones futuras, aspirando a una comprensión y evaluación más matizada y eficaz de los resúmenes automáticos.

2. Estado del Arte

En la última década, el procesamiento del lenguaje natural ha experimentado avances significativos, siendo la generación de resúmenes automáticos uno de los campos centrales de estas innovaciones. Esta tarea desafiante busca transformar textos extensos en representaciones más breves, conservando la esencia y la información clave del contenido original. La relevancia de este campo no solo radica en la creciente necesidad de gestionar y sintetizar la sobrecarga de información en nuestra era digital, sino también en las complejas técnicas y enfoques desarrollados para abordar el problema. Esta sección delimita el panorama actual de los modelos y técnicas predominantes en la generación de resúmenes automáticos, poniendo de relieve sus avances, limitaciones y el papel fundamental del feedback humano en este ecosistema.

2.1. Modelos de resúmenes automáticos

La creación de resúmenes automáticos representa un pilar en el procesamiento del lenguaje natural, enfocándose en destilar información extensa en presentaciones más breves y manejables. Diferentes enfoques han surgido a lo largo de los años para abordar esta tarea, reflejando la evolución de la tecnología y los métodos de investigación. En esta sección, se abordarán tres enfoques principales que han caracterizado esta evolución: modelos extractivos, abstractivos y aquellos basados en aprendizaje por refuerzo.

2.1.1. Modelos extractivos

Los modelos extractivos de resumen automático se centran en identificar y extraer fragmentos clave o segmentos directamente del documento fuente para formar el resumen. En lugar de reescribir o parafrasear el contenido, estos modelos seleccionan partes del texto original que se consideran más informativas o relevantes. A lo largo del tiempo, se han propuesto diversos métodos para esta estrategia de resumen, desde enfoques basados en heurísticas, como el presentado por Dorr et al. 2003, hasta técnicas más modernas y sofisticadas basadas en redes neuronales, como SummaRuNNer de Nallapati et al. 2017. A pesar de la simplicidad relativa de los modelos extractivos en comparación con los métodos abstractivos, han demostrado ser efectivos y ofrecer resultados competitivos en muchos conjuntos de datos.

[1]La escala Likert es una escala psicométrica utilizada en cuestionarios para representar el nivel de acuerdo o desacuerdo del encuestado con una afirmación particular. Consiste típicamente en un conjunto de declaraciones y el encuestado debe indicar su nivel de acuerdo con cada una, usualmente en una escala que va desde "totalmente en desacuerdo" hasta "totalmente de acuerdo". Más información disponible en el siguiente enlace.

2.1.2. Modelos abstractivos

El resumen automático abstractivo es una tarea de procesamiento del lenguaje natural (NLP) que busca generar un resumen conciso y coherente de un texto original, utilizando palabras y frases que no necesariamente aparecen en el documento fuente. A diferencia del resumen extractivo, que simplemente selecciona y reorganiza partes del texto original, el resumen abstractivo reformula y sintetiza el contenido para crear un resumen que, a menudo, parece haber sido escrito por un humano.

A lo largo de los años, se han realizado numerosas investigaciones para mejorar y refinar los métodos de resumen abstractivo. El campo ha evolucionado desde técnicas basadas en reglas y heurísticas, como las presentadas en Dorr et al. 2003, hasta enfoques más recientes que aprovechan el poder del aprendizaje profundo, como se muestra en Chopra et al. 2016 y Rush et al. 2015. Un desarrollo particularmente notable en este espacio es el trabajo realizado en See et al. 2017, que propone una arquitectura que aborda específicamente los problemas de reproducción inexacta de detalles y repetición en los resúmenes generados.

Estos avances en técnicas y metodologías subrayan la importancia continua y el interés en la capacidad de generar resúmenes que no sólo sean informativos y concisos, sino que también reflejen con precisión el mensaje y la intención del texto original. Sin embargo, dada la naturaleza subjetiva de la calidad de un resumen, todavía hay retos en la evaluación y cuantificación de la calidad sin la intervención humana.

2.1.3. Aprendizaje por refuerzo

El aprendizaje por refuerzo (RL) es un paradigma en el cual un agente aprende cómo actuar en un entorno al recibir retroalimentación en forma de recompensas o penalizaciones basadas en sus acciones. En el contexto de la generación de resúmenes automáticos, RL ofrece un enfoque innovador: en lugar de simplemente predecir las oraciones o fragmentos más relevantes basados en etiquetas previas, los modelos son entrenados para generar resúmenes y luego recibir retroalimentación según la calidad de esos resúmenes, generalmente basándose en métricas como ROUGE.

Los enfoques tradicionales generación de resúmenes automáticos, que se basan en modelos supervisados, a menudo dependen de etiquetas heurísticas o conjuntos de datos predefinidos. El RL, en cambio, permite que los modelos exploren una variedad más amplia de resúmenes y ajusten sus estrategias basándose directamente en la calidad del resumen producido. Esto es particularmente útil para resumir textos donde no hay un único "correcto" resumen, permitiendo que el modelo explore y aprenda de múltiples soluciones posibles.

Varios trabajos, incluyendo Narayan et al. 2018 y Dong et al. 2018, han enfocado sus esfuerzos en desarrollar modelos que pueden seleccionar y rankear oraciones de una manera que maximice estas recompensas, superando a las técnicas tradicionales basadas en reglas.

Adicionalmente, Wu et al. 2018 resalta la importancia de la coherencia en los resúmenes generados, introduciendo un modelo que puede capturar de forma efectiva las relaciones semánticas y sintácticas entre oraciones, resultando en resúmenes más coherentes y legibles.

En conjunto, estas investigaciones indican una trayectoria prometedora para el uso del aprendizaje por refuerzo en la generación de resúmenes, presentando soluciones que se ajustan más fielmente a las métricas deseadas y generando resúmenes más informativos y coherentes.

2.2. Desafíos en la Generación Automática de Resúmenes

El empleo de LLMs pre-entrenados ha ganado popularidad, pues ha potenciado el rendimiento en múltiples tareas del procesamiento del lenguaje natural. Al adaptar estos modelos a tareas específicas, como la generación de resúmenes, se hace un "fine-tuning" mediante aprendizaje supervisado, orientado a maximizar la verosimilitud de textos anotados por humanos. No obstante, si bien estos métodos han demostrado ser eficientes, como se discute en la Sección 2.1.2, todavía persiste una falta de alineamiento entre las funciones objetivo y las preferencias humanas en la generación de resúmenes de calidad, como se destaca en Stiennon et al. 2020.

Esta falta de alineamiento se debe a numerosas causas, entre las cuales cabe destacar las siguientes:

- **"Alucinación" de contenido:** Maynez et al. 2020 señalan un fenómeno preocupante en el que los modelos de generación de resúmenes abstractivos tienden a "alucinar", o introducir, contenido que no está presente en el texto fuente original. A pesar de que los modelos pre-entrenados han demostrado ser más precisos en la generación de resúmenes, aún no están exentos de este problema. Además, esta investigación destaca las limitaciones de métricas tradicionales como ROUGE en la evaluación de resúmenes automáticos y sugiere que las métricas basadas en inferencia semántica podrían ser indicadores más fiables de la calidad de un resumen.
- **Aprendizaje secuencial y dependencias:** El trabajo de Ross et al. 2010 profundiza en el reto del aprendizaje secuencial, especialmente en el aprendizaje por imitación. En tales escenarios, las predicciones o acciones futuras dependen de decisiones anteriores, rompiendo así las asunciones típicas de independencia y distribución idéntica (i.i.d.) que son fundamentales en muchas técnicas de aprendizaje automático. Estas dependencias pueden resultar en modelos que no se desempeñan de manera óptima en la generación de texto.
- **Sesgo de exposición y generalización:** El sesgo de exposición (Schmidt 2019), es un desafío clave en modelos generativos autoregresivos. La raíz del problema es que, durante el entrenamiento, los modelos se exponen a datos reales, pero durante la generación o prueba, a menudo generan y se basan en sus propios contextos previos. Esto lleva a un falta de alineamiento entre lo que el modelo ha "visto" durante el entrenamiento y lo que produce durante la generación, afectando su capacidad para generalizar correctamente en contextos no vistos.
- **Equilibrio entre diversidad y calidad:** H. Zhang et al. 2020 abordan el dilema entre mantener la diversidad y asegurar la calidad en la generación de lenguaje. La elección de cómo decodificar, o cómo un modelo decide las próximas palabras o frases a generar, juega un papel fundamental en la calidad final del texto. Su trabajo destaca cómo diferentes técnicas pueden tener un rendimiento similar cuando se prioriza la diversidad, pero algunas, como el "muestreo de núcleo" o "muestreo Top-p" (Holtzman et al. 2019), son superiores cuando se prioriza la calidad.
- **"Degeneración" del texto en modelos avanzados:** Por último, el estudio de Holtzman et al. 2019 arroja luz sobre un problema común en modelos de lenguaje neuronal profundo: la "degeneración" del texto. A pesar de su capacidad para entender el lenguaje, estos modelos, cuando se utilizan para la generación, a menudo producen salidas que son repetitivas y carecen de diversidad y originalidad.

A pesar de los avances en el campo de los modelos de lenguaje, las investigaciones mencionadas subrayan que todavía hay desafíos significativos por resolver. La clave para superar estas barreras puede residir en la interacción constante y el feedback entre humanos y máquinas, asegurando así que los sistemas generativos satisfagan de manera óptima las necesidades y expectativas humanas.

2.3. Evaluación y limitaciones de las métricas Automáticas

En el ámbito de la generación automática de resúmenes, uno de los desafíos más persistentes y discutidos es la evaluación de la calidad de los resúmenes generados. A pesar de la rapidez y consistencia de las métricas automáticas, a menudo no reflejan adecuadamente los juicios humanos, considerados el estándar de oro en la evaluación. Un estudio relevante (Novikova et al. 2017) demostró que métricas populares como BLEU (Papineni et al. 2002) y ROUGE (Lin 2004) tienen correlaciones débiles con las valoraciones humanas, especialmente en sistemas de generación automática de lenguaje mediante modelos supervisados. Además, se resaltó que la efectividad de estas métricas varía según el sistema y el conjunto de datos utilizado.

ROUGE (Lin 2004), ampliamente reconocido en el área, ha recibido también críticas. Según Schluter 2017, aunque esté optimizado, no garantiza la producción de resúmenes de alta calidad. Otras investigaciones (Paulus et al. 2017) respaldan esta visión, subrayando el riesgo de que ROUGE favorezca resúmenes menos legibles.

La tensión entre métricas automáticas y juicio humano es más evidente en Chaganty et al. 2018. Este trabajo sugiere que depender exclusivamente de métricas automáticas puede ser engañoso y llevar a pasar por alto avances significativos en los modelos de resumen. Por su parte, Kryscinski et al. 2019 ofrece una perspectiva holística, identificando fallos en métricas, conjuntos de datos y los modelos de resumen en sí.

Sin embargo, no es el único estudio que ha cuestionado las métricas actuales. Los autores Fabbri et al. 2020 llevaron a cabo una re-evaluación exhaustiva de las métricas empleadas en la generación automática de resúmenes. Para esto, emplearon el corpus CNN/DailyMail (Hermann et al. 2015), un conjunto de datos originalmente diseñado para tareas de QA pero que posteriormente fue adaptado para la generación automática de resúmenes por Nallapati et al. 2016.

En dicho estudio, no se limitaron a la métrica ROUGE (Lin 2004), sino que evaluaron también otras métricas reconocidas como BLEU (Papineni et al. 2002), METEOR (Lavie et al. 2007) y BERTScore (T. Zhang et al. 2019), entre otras.

De esta investigación, se destacan las siguientes conclusiones:

- Existe una variabilidad significativa en la correlación de diferentes métricas con el juicio humano.
- No hay una métrica que se correlacione de manera perfecta con las evaluaciones humanas. Esto sugiere la necesidad de una combinación de métricas para lograr una evaluación más completa.
- Aunque algunas métricas contemporáneas superan a las tradicionales, como ROUGE, 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.

Para concluir, a medida que los modelos de inteligencia artificial evolucionan y se vuelven más sofisticados, crece el consenso en la comunidad científica sobre la necesidad urgente de avanzar en la investigación y mejorar las métricas actuales. Es fundamental adaptar y evolucionar estos métodos teniendo en cuenta la complejidad y diversidad de los textos, así como las expectativas del usuario final. Este panorama refuerza el llamado a la comunidad investigadora a perfeccionar no solo los modelos, sino también las métricas empleadas para evaluar su eficacia.

2.4. Feedback humano como piedra angular

La generación automática de resúmenes, a pesar de los avances en modelos de lenguaje, todavía enfrenta desafíos significativos, como se discutió en la Sección 4.3. La problemática central radica en que las métricas automáticas tradicionales, como ROUGE, no siempre reflejan adecuadamente las verdaderas expectativas humanas, lo cual pone en cuestión su utilidad real para evaluar la calidad del resumen 2.3. Esta brecha entre las métricas automáticas y el juicio humano ha enfocado la atención hacia la importancia del "feedback humano" en el proceso.

El estudio de Böhm et al. 2019 explora precisamente cómo mejorar la generación de resúmenes mediante Aprendizaje por Refuerzo (RL) incorporando este feedback. A diferencia de las técnicas tradicionales de RL que dependen de métricas como ROUGE, este trabajo propone una función de recompensa basada en las valoraciones humanas. Después de evaluar 2,500 resúmenes, se presenta una función que opera sin necesidad de resúmenes de referencia, mostrando una correlación superior con las valoraciones humanas en comparación con ROUGE y METEOR. Además, es notable que los sistemas desarrollados con esta función tienen mejor desempeño que otros sistemas avanzados, logrando tal proeza con menos datos de entrenamiento.

Por su parte, Ziegler et al. 2019 ajustaron modelos de lenguaje mediante RL para alinearlos con las preferencias humanas. Aunque se obtuvieron buenos resultados en tareas como la continuación de texto, en resumen, los modelos tendieron a ser principalmente extractivos. El énfasis de la investigación recae en la relevancia de las recompensas basadas en feedback humano y las limitaciones inherentes de las métricas convencionales.

Asimismo, Leike et al. 2018 abordan cómo alinear agentes de RL con las intenciones del usuario, presentando el concepto de "modelo de recompensa". Este modelo se concentra en aprender una función de recompensa a partir de la interacción directa con el usuario, y el estudio destaca los desafíos en la implementación de este enfoque en dominios más complejos.

Para concluir, el estudio sobre el aprendizaje de resúmenes basado en feedback humano (Stiennon et al. 2020), que ha sido la principal referencia en nuestra investigación, subraya las limitaciones tanto en la fase de entrenamiento de este tipo de modelos, como su misma evaluación. Dicha investigación

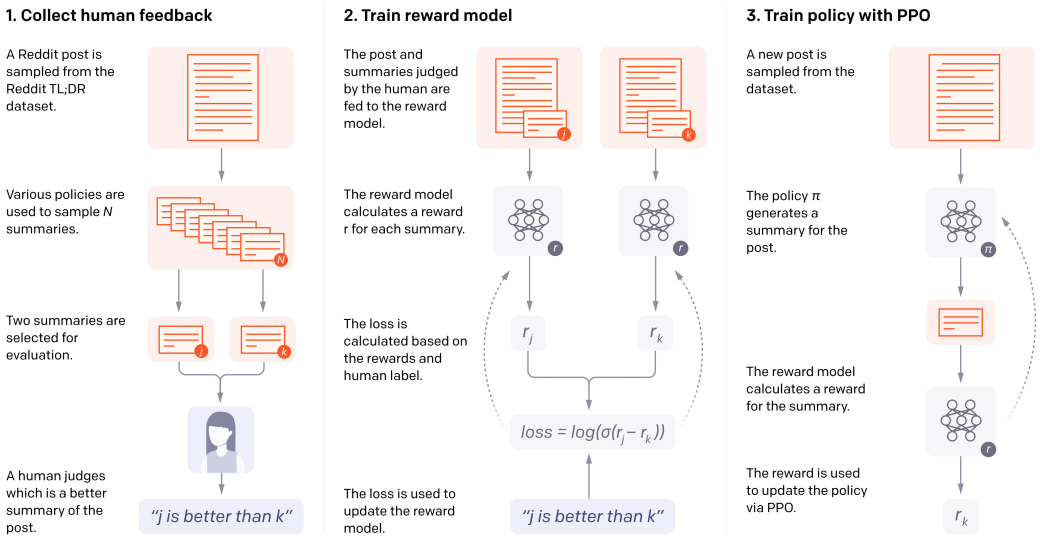

Figura 1: Diagrama del proceso de recolección de datos con "feedback humano", entrenamiento de modelos de recompensa y entrenamiento de "policies" mostrado en Stiennon et al. 2020.

propone entrenar modelos con un enfoque en las preferencias humanas, utilizando estas mismas para definir una función de recompensa en el RL. El proceso detallado en esta propuesta (ilustrado en la Figura 1) se fundamenta en una metodología estructurada de tres pasos que se repiten de forma iterativa.

- **Paso 1:** Recolección y comparación de muestras, donde resúmenes de distintas fuentes son evaluados por expertos humanos.
- **Paso 2:** Entrenamiento del "modelo de recompensa", centrado en predecir las preferencias humanas basadas en el feedback proporcionado.
- **Paso 3:** Optimización de una "política" en función del "modelo de recompensa". El output del modelo de recompensa se utiliza como "recompensa" en la optimización del sistema de aprendizaje por refuerzo, utilizando el algoritmo PPO (Proximal Policy Optimization (Schulman et al. 2017)).

Este enfoque, aunque lineal en su presentación, se adapta y evoluciona en la práctica con la acumulación continua de datos y etiquetas. Sorprendentemente, los resultados revelan que estos modelos superan en desempeño a los resúmenes de referencia humanos y a otros modelos formados mediante aprendizaje supervisado, demostrando su aplicabilidad en diversos conjuntos de datos. Estos hallazgos sugieren una tendencia hacia el futuro, donde el feedback humano jugará un papel esencial en el desarrollo y perfeccionamiento de sistemas automáticos.

3. Metodología y diseño experimental

En los capítulos anteriores, hemos discutido el contexto y los objetivos de nuestro estudio. Ahora, nos centraremos en el corazón de nuestra investigación: la metodología y el diseño experimental (ver Sección 3.2 y Sección 3.3).

Primero, en la Sección 3.1 de "Metodología general", presentaremos una visión general de nuestro enfoque metodológico, proporcionando una estructura holística de cómo se ha diseñado el estudio para abordar nuestras preguntas de investigación.

A continuación, en la Sección 3.2 de "Configuración experimental", detallaremos los aspectos prácticos de cómo se llevó a cabo la investigación. En la Sección 3.2.1 se explicará la procedencia del dataset, la estructura del mismo, y los diversos modelos y enfoques empleados para su generación. Esto incluirá también cómo se prepararon los datos para el análisis mediante ingeniería de características y pre-procesado (Sección 3.2.2), y qué tecnologías y herramientas se emplearon

para respaldar y optimizar cada fase de la experimentación (Sección 3.2.3). Por último se definirá en detalle la metodología propuesta para identificar características predictivas de la calidad de un resumen (Sección 3.2.4).

En la Sección 3.3 de "Construcción de experimentos", describiremos cada uno de los experimentos que realizamos en profundidad. Para cada experimento, proporcionaremos información sobre cómo se llevó a cabo, qué descubrimos y cómo estos hallazgos se relacionan con nuestras preguntas de investigación (Sección 3.3.1, Sección 3.3.2, Sección 3.3.3).

Finalmente, concluiremos este capítulo con una "Discusión general" (Sección 3.4), donde integraremos y analizaremos los resultados de todos nuestros experimentos en conjunto. En esta sección, evaluaremos cómo nuestros hallazgos contribuyen a nuestra comprensión del tema en cuestión y discutiremos las implicaciones de nuestro trabajo.

En conjunto, este capítulo proporcionará una visión completa de cómo se llevó a cabo nuestra investigación, desde la concepción y diseño hasta la realización y análisis.

3.1. METODOLOGÍA GENERAL

El núcleo de este estudio radica en desentrañar una respuesta contundente a la siguiente pregunta: ¿En el contexto actual, qué métricas o rasgos son determinantes para evaluar efectivamente la calidad de los resúmenes? Enfocándonos en este interrogante, se toma como referencia un dataset creado por OpenAI[2], cuyos detalles se abordarán más adelante. Este dataset ha sido anotado de dos formas distintas. Para nuestro propósito, se pondrá énfasis en la comparación de resúmenes, donde cada observación consiste en un texto de referencia, dos resúmenes (ya sean generados automáticamente o elaborados por anotadores humanos) y la elección del anotador sobre cuál de los dos resúmenes representa mejor al texto original. Esta decisión es binaria y se refleja en la estructura que se describirá a continuación.

Input {Texto}	Resumen A	Resumen B	Target {A, B}

Conociendo nuestro punto de partida y teniendo claro el conjunto de datos con el que trabajamos, planteamos un problema de aprendizaje automático. Este tiene como objetivo modelar la decisión tomada por los anotadores, utilizando métricas y rasgos intrínsecos de los resúmenes como predictores. Una vez definido y entrenado este modelo, lo emplearemos para determinar las métricas y rasgos que tienen una mayor influencia en la toma de decisiones, y por ende, aquellos que se correlacionan con la percepción de calidad en un resumen.

A fin de reforzar la validez de nuestros hallazgos, también realizaremos una serie de experimentos complementarios que justifiquen las asunciones adoptadas. Por último, abordaremos una exploración detallada sobre cómo el poder predictivo de las métricas y rasgos fluctúa en función de la calidad intrínseca del resumen en cuestión, y si esto tiene una repercusión significativa en la ponderación de dichas métricas.

3.2. CONFIGURACIÓN EXPERIMENTAL

La configuración experimental es una pieza clave en este trabajo, ya que establece las bases sobre las cuales se desarrollan y evalúan los métodos propuestos. En esta sección, nos sumergimos en la metodología detrás de la recolección y procesamiento de los datos utilizados en este estudio. Describimos, con detalle, los pasos iniciales que involucran la selección y limpieza de datos, así como la ingeniería de características que potenciará nuestros modelos. Además, se introducen las tecnologías y herramientas que respaldan y optimizan cada etapa de nuestra experimentación. Posteriormente, se presenta `ExplainSumm`, una metodología novedosa diseñada para abordar las preguntas de investigación propuestas. Esta sección se estructura en varios subapartados, desde la recolección de datos hasta la descripción detallada de la nueva metodología propuesta, ofreciendo una visión completa y profunda del proceso experimental.

[2]En el siguiente enlace se puede encontrar ejemplos de este corpus.

3.2.1. RECOLECCIÓN DE DATOS

En esta sección se detallan el proceso y las herramientas utilizadas para la recolección y tratamiento de datos, según lo expuesto en Stiennon et al. 2020. Se explica la procedencia de los conjuntos de datos, los criterios de valoración y evaluación de los resúmenes y los diversos modelos y enfoques empleados para su generación. Además, se aborda la relevancia del feedback humano y su integración en el proceso.

Dataset. El conjunto de datos utilizado en este estudio es parte del proyecto "Learning to summarize with human feedback" (Stiennon et al. 2020) de OpenAI. Se puede acceder y explorar estos datos directamente en la página web creada para el proyecto. Estos datos se derivan inicialmente del dataset TL;DR (Völske et al. 2017), un corpus que engloba aproximadamente 3 millones de publicaciones de *reddit.com*, abarcando una diversidad de temas (*subreddits*). Además de las publicaciones, el conjunto de datos también incluye resúmenes de las entradas escritos por sus propios autores, conocidos como TL;DRs. Este conjunto de datos se ha enriquecido con resúmenes de los posts realizados por anotadores humanos. Se han seleccionado exclusivamente aquellos posts cuyos resúmenes de referencia (escritos por anotadores humanos) tienen un tamaño de entre 24 y 48 tokens para minimizar sesgos relacionados con la longitud de los resúmenes, resultando en una selección más concentrada de aproximadamente 120,000 publicaciones. Dicho lo cual, el conjunto de datos que se tiene hasta el momento es el siguiente:

Reddit	Sub-reddit	Título	Texto	Resumen de Ref.

La elección del dataset TL;DR sobre el más reconocido dataset CNN/DM (Nallapati et al. 2016) se justifica porque, en este último, modelos de referencia de tipo extractivo como *lead-3 model* superan en rendimiento a los resúmenes escritos por humanos, situación que no se presenta en TL;DR. Esta observación es detalladamente expuesta en el Apéndice E de la mencionada publicación.

Directrices de evaluación. Según los autores (Stiennon et al. 2020) se evalúa la calidad del resumen por cuán fielmente transmite el post original a un lector que solo puede leer el resumen y no el post.

Generación de resúmenes automáticos. A partir de la información detallada anteriormente, se generan resúmenes automáticos utilizando diferentes algoritmos para cada post del conjunto de datos mencionado. Posteriormente, estos resúmenes son evaluados por un anotador humano mediante diversos criterios. En cuanto a los modelos de generación automática de resúmenes, estos pueden clasificarse en tres categorías principales:

1. **Pre-trained models:** Son versiones de GPT-3 que no han sido ajustadas (fine-tuned) para una tarea específica.
2. **Supervised baselines:** Modelos que han sido ajustados (fine-tuned) para predecir TL;DRs (resúmenes de las entradas escritos por los propios autores del post).
3. **Human feedback models:** Modelos que, además de lo anterior, han sido adicionalmente ajustados (fine-tuned) utilizando un conjunto de datos de aproximadamente 65K comparaciones de resúmenes. El proceso detallado de este tipo de modelos se puede apreciar en la Figura 1.

Por lo tanto, estos modelos representan diferentes etapas o enfoques de entrenamiento, desde modelos pre-entrenados generales hasta modelos ajustados específicamente para tareas de resumen, ya sea a través de supervisión directa o retroalimentación humana. Además, para cada uno de estos modelos, existen diferentes configuraciones de parámetros y "políticas".

Incorporación del feedback humano. En este punto, tenemos para cada post de reddit diferentes resúmenes generados por diferentes algoritmos de generación automática, y el resumen de referencia escrito por un anotador humano. A partir de aquí, el dataset se anota de formas diferentes:

1. **Comparación de resúmenes (comparisons):** Dado un texto, los anotadores confrontan dos resúmenes proporcionados y eligen el que consideran más fiel en aspectos como esencia, claridad, precisión, propósito, concisión y estilo. También asignan un valor, dentro de

una escala de 9 puntos, reflejando su grado de confianza en que un resumen es superior al otro. En los resúmenes proporcionados, también se incluyen resúmenes de referencia escritos por evaluadores humanos.

2. **Evaluación de calidad de los resúmenes (axis):** Los anotadores juzgan la calidad de los resúmenes bajo cuatro dimensiones, usando una escala de 7 puntos. Estas dimensiones son:
   - cobertura de la información,
   - precisión en la representación del contenido original,
   - coherencia del resumen,
   - y valoración general del resumen.

A continuación, se muestra la estructura del dataset de "comparisons":

Reddit	Sub-reddit	Título	Texto	Resumen A	Resumen B	Elección {A, B}

Este trabajo se fundamenta en ambas anotaciones para discernir los patrones distintivos de un resumen automático de calidad, y cómo estos se alinean con las inclinaciones de los anotadores, pero sin duda el dataset "comparisons" es el corpus de referencia que se utiliza en este trabajo, y su punto de partida.

3.2.2. Pre-procesamiento y Feature Engineering

La preparación de datos es una etapa crucial en cualquier proyecto de aprendizaje automático, especialmente cuando se trabaja con datos de texto. Nuestro proceso se divide en dos etapas: pre-procesamiento y feature engineering. Estas etapas están diseñadas para preparar y enriquecer nuestro conjunto de datos para su posterior análisis.

Pre-procesamiento. Asegurar que el dataset sea coherente, esté libre de ruido y en un formato adecuado para las etapas posteriores del análisis es esencial. Para lograr esto, hemos desarrollado una clase llamada `Preprocessor`[3].

La principal función de la clase `Preprocessor` es la limpieza y transformación inicial del conjunto de datos, garantizando que esté listo para el análisis y la extracción de características.

Los métodos principales de la clase se encargan de:

- Eliminar espacios blancos innecesarios en las columnas de texto.
- Transformar ciertos tipos de datos, como la conversión de columnas de tipo "object" con valores booleanos a columnas de tipo booleanas "object".
- Hacer reemplazos basados en patrones, por ejemplo, eliminar números que se han introducido como anotaciones adicionales.
- Gestión de valores nulos, garantizando que las columnas clave no contengan valores nulos o vacíos.

Feature Engineering. Tras el pre-procesamiento, el conjunto de datos es sometido a una fase en la que se realizan técnicas de ingeniería de características. Esta etapa está gestionada por la clase `FeatureEngineering`[4].

El propósito principal de esta clase es enriquecer el conjunto de datos mediante el cálculo de diversas métricas y rasgos que podrían ser cruciales para identificar patrones y características de resúmenes de calidad. La clase se centra en las siguientes categorías principales de métricas y rasgos:

1. **Métricas de Evaluación de Resúmenes (Resumen vs. Resumen de Referencia)**
   - Se utilizan medidas conocidas como las puntuaciones **ROUGE** para evaluar la similitud entre resúmenes y resúmenes de referencia. Estas puntuaciones, como **ROUGE-1**, **ROUGE-2**

[3] El código de la clase `Preprocessor` se encuentra en el siguiente enlace del proyecto.

[4] El código de la clase `FeatureEngineering` se encuentra en el siguiente enlace del proyecto.

y **ROUGE-L**, permiten entender aspectos como el solapamiento de tokens y n-gramas (secuencias contiguas de n palabras) entre ambos textos.
   - De igual forma, se emplea la métrica **BLEU**, que evalúa la calidad de los resúmenes generados en comparación con los resúmenes de referencia, considerando la coincidencia de palabras y frases.
   - Adicionalmente, se introduce una nueva perspectiva de evaluación utilizando modelos de lenguaje pre-entrenados diseñados para generar embeddings (representaciones multidimensionales de texto) de oraciones y texto. En este caso, embeddings de resúmenes.
   - Estas métricas proporcionan un enfoque más centrado en el contenido que en la coincidencia exacta de palabras. Los embeddings proyectan el texto en un espacio vectorial denso de altas dimensiones, y estos pueden ser comparados mediante la **Similitud del Coseno**, para medir así el grado de afinidad entre resúmenes.

2. **Métricas de Correspondencia Textual (Texto Original vs. Resumen):**
   - Dentro entre grupo de métricas, al igual que en el apartado anterior, se calcula la **Similitud del Coseno** entre el embedding del texto original, y el embedding del resumen dado.
   - Otra métrica que ha sido implementada es el **Ratio de Compresión**, que indica cuánto ha sido condensado el texto original para producir el resumen. Este ratio se obtiene dividiendo la longitud del resumen entre la longitud del texto original.
   - Por último, la **Similitud de Jaccard** evalúa el solapamiento de palabras o n-gramas entre el texto original y el resumen, calculándose tanto para n-gramas de tamaño 1 como para tamaño 2. Una Similitud de Jaccard más alta indica un mayor grado de palabras compartidas entre el texto de entrada y el resumen.
3. **Rasgos de Legibilidad y Complejidad Lingüística:** Esta categoría se encarga de analizar cómo se presenta el resumen en términos de accesibilidad y complejidad lingüística. Se centra en obtener y calcular las siguientes características:
   - El **Recuento de Sílabas**, **Léxicos**, **Oraciones**, **Caracteres**, **Letras**, **Polisílabos** y **Monosílabos** aportan información detallada sobre la estructura y construcción del resumen.
   - Usando estos rasgos lingüísticos, se calcula el **Flesch Reading Ease Score (RE)**, una medida cuantitativa de la facilidad de lectura del texto. Este índice ofrece una perspectiva sobre cuán accesible y legible es el resumen para el lector final.

Respecto a la generación de embeddings, se utiliza la librería `SentenceTransformers`[5] en Python, un framework especializado en la generación de embeddings de oraciones, texto e imágenes, y que se describe detalladamente en el artículo "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks" de Reimers et al. 2019. En concreto, para proyectar estos embeddings, se ha seleccionado el modelo `all-MiniLM-L6-v2`. Este modelo tiene la capacidad de mapear oraciones y párrafos a un espacio vectorial denso de 384 dimensiones, lo que lo hace particularmente útil para tareas como la búsqueda o agrupación semántica.

Para una explicación más detallada de las métricas y rasgos propuestos, se puede consultar el apéndice A.

3.2.3. Framework

Este trabajo emplea diversas herramientas y técnicas que no solo facilitan y optimizan el proceso de modelado, sino que también justifican las decisiones tomadas. Estas tecnologías contribuyen a la eficiencia, precisión y transparencia de los modelos, así como a la justificación de las predicciones. A continuación, se describen algunas de las tecnologías clave empleadas.

Extreme Gradient Boosting. XGBoost[6], es un algoritmo de aprendizaje supervisado basado en el principio de *boosting*. Optimiza el árbol de decisión de gradiente para ser más eficiente y preciso.

La esencia de XGBoost radica en construir un conjunto de árboles de decisión de manera secuencial, donde cada árbol rectifica los errores del anterior. Formalmente, si tenemos un conjunto de datos $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ con $x_i \in \mathbb{R}^m$ y $y_i \in \mathbb{R}$, el modelo predictivo es:

[5]La documentación de `SentenceTransformers` se encuentra el siguiente enlace.

[6]La documentación de `XGBoost` se encuentra en el siguiente enlace.

$$f(x) = \sum_{k=1}^{K} T_k(x)$$

donde $T_k(x)$ es el k-ésimo árbol de decisión. La función objetivo para optimizar en XGBoost es:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{K} \Omega(T_k)$$

donde l es una función de pérdida diferenciable que mide la discrepancia entre la predicción $\hat{y}_i^{(t)}$ y el valor verdadero y_i, y $\Omega(T_k)$ es una función de regularización que penaliza la complejidad del árbol T_k.

Un atributo distintivo de XGBoost es su habilidad para entrenarse usando una GPU (Graphics Processing Unit), la cual, debido a su arquitectura paralela, es altamente eficiente para operaciones matriciales y vectoriales, comunes en el entrenamiento de modelos de aprendizaje automático. Así, XGBoost acelera el proceso de entrenamiento al compararlo con el entrenamiento exclusivo en CPU.

Optuna. Conocido por ser una potente herramienta en el ámbito del aprendizaje automático, Optuna[7] es una librería de software de código abierto dedicada a la optimización automática de hiperparámetros. El principal objetivo de esta herramienta es ofrecer una optimización que sea eficiente tanto en tiempo de ejecución como en calidad de resultados.

El enfoque de Optuna se basa en la interacción entre estudio y trial:

- **Estudio**: Corresponde a una única optimización de hiperparámetros centrada en una función objetivo específica.
- **Trial**: Se refiere a cada ejecución individual de la función objetivo, llevada a cabo con un conjunto determinado de hiperparámetros.

La función objetivo, que en este trabajo es la minimización del error medio en el conjunto test de la validación cruzada[8], es la que guía este proceso de optimización. Su tarea es retornar un valor escalar que Optuna se esforzará por minimizar. Para seleccionar hiperparámetros, Optuna recurre a diversos métodos de muestreo, entre ellos:

- **Random Sampling**: Realiza un muestreo aleatorio de hiperparámetros.
- **TPE (Tree-structured Parzen Estimator)**: Se basa en un algoritmo que estima la densidad del modelo para el muestreo de hiperparámetros.
- **Grid Sampling**: Consiste en un muestreo exhaustivo a partir de un conjunto de valores previamente definidos.
- **CMA-ES (Covariance Matrix Adaptation Evolution Strategy)**: Usa un algoritmo de optimización que se fundamenta en estrategias evolutivas.

Una de las características que distingue a Optuna es el *pruning*, que tiene como finalidad interrumpir aquellas iteraciones que no muestren un rendimiento prometedor. Este proceso permite ahorrar recursos computacionales al evitar la ejecución de iteraciones que, según resultados intermedios, tienen pocas probabilidades de superar a otros ya finalizados.

Optuna presenta una integración sencilla con diversas bibliotecas ampliamente utilizadas en aprendizaje automático, como lo son TensorFlow, PyTorch y Scikit-learn. Además, ofrece herramientas de visualización que facilitan el análisis y comprensión de los resultados obtenidos en la optimización.

[7] La página web de este Framework de optimización se encuentra en el siguiente enlace.

[8] La función objetivo "OptunaObjective" se define en el siguiente enlace al repositorio GitHub del proyecto.

SHAP (Shapley Additive exPlanations). SHAP[9] utiliza la teoría de valores de Shapley, derivada de la teoría de juegos, para explicar las salidas de cualquier modelo de aprendizaje automático. Proporciona una interpretación unificada y coherente de las predicciones del modelo, descomponiéndolas en función de la contribución de cada característica.

La teoría de valores de Shapley proviene de la teoría de juegos y se usa para determinar cómo distribuir un valor de juego total entre varios jugadores, dependiendo de su contribución al juego. En el contexto de SHAP, cada característica de un modelo se considera como un "jugador" y la "recompensa" es la predicción del modelo.

Para una predicción dada, el valor SHAP de una característica es el promedio ponderado de sus márgenes contributivos, tomados en cuenta todos los posibles conjuntos de características. Los valores SHAP tienen las siguientes propiedades:

- **Localidad**: Cada predicción tiene su propio conjunto de valores SHAP, permitiendo explicaciones detalladas para predicciones individuales.
- **Aditividad**: La suma de los valores SHAP para todas las características es igual a la diferencia entre la predicción del modelo y la salida promedio del modelo.
- **Consistencia**: Si el modelo cambia de manera que una característica tiene un impacto mayor en la predicción, su valor SHAP no disminuye.

SHAP es extremadamente útil para:

- Comprender la importancia de características en modelos complejos.
- Diagnosticar posibles problemas o sesgos en un modelo.
- Explicar las predicciones individuales a los interesados.

SHAP es compatible con muchas de las librerías más populares de aprendizaje automático, facilitando la explicación de modelos que van desde árboles de decisión hasta redes neuronales. Por estas razones, en este trabajo se utilizará la librería SHAP con el modelo predictivo XGBoost para interpretar las decisiones de los anotadores.

Validación Cruzada (Cross-Validation). En este trabajo, se ha optado por utilizar la técnica de **K-fold cross-validation**. Es importante destacar que, al realizar la segmentación, se ha tenido especial cuidado en garantizar que cada subconjunto contenga "posts" de Reddit únicos. Esta medida se adoptó para prevenir problemas de **Data Leakage**, asegurando que la información de un mismo "post" no esté en los conjuntos de entrenamiento y validación simultáneamente. Para llevar a cabo la validación cruzada K-fold, no se han utilizado bibliotecas de terceros específicas. En su lugar, se desarrolló un código personalizado[10] para este propósito, utilizando la librería Numpy. Este enfoque personalizado ofreció un control más granular sobre el proceso de segmentación y garantizó una correcta aplicación de la técnica.

La validación cruzada es una técnica empleada para evaluar la capacidad de un modelo para generalizar en un conjunto de datos. Es útil para identificar problemas como el sobreajuste y para seleccionar el modelo óptimo de varias alternativas. La estrategia de la validación cruzada involucra segmentar el conjunto de datos original en k subconjuntos (o "folds") de tamaño similar. El modelo se entrena con $k-1$ subconjuntos y se valida con el subconjunto restante, repitiendo el proceso k veces y rotando el subconjunto de validación. Después, se promedian los resultados para obtener una única medida de rendimiento.

Matemáticamente, un conjunto de datos $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ se divide en k subconjuntos disjuntos $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$. Para cada $i \in \{1, 2, \ldots, k\}$, se entrena el modelo con $\mathcal{D} \setminus \mathcal{D}_i$ y se valida con $\mathcal{D}_i$.

Existen varios tipos de Validación Cruzada:

- **K-fold cross-validation**: Es el método estándar descrito anteriormente donde el conjunto de datos se divide en k subconjuntos.

[9] En el siguiente enlace se muestra la documentación técnica de la librería python SHAP.

[10] Tarea definida en la función "text_folds" del siguiente enlance al repositorio GitHub.

- **Leave-One-Out (LOO)**: Es un caso especial de validación cruzada donde $k = n$, es decir, en cada iteración se deja un solo dato para validación y el resto se utiliza para el entrenamiento.
- **Stratified K-fold**: Es una variación del K-fold donde se garantiza que cada fold tenga aproximadamente la misma proporción de muestras de cada clase objetivo como el conjunto completo.
- **Time Series Cross-Validation**: Especialmente útil cuando se trabaja con series temporales, garantizando que el orden temporal se mantenga al dividir los datos.

Las principales ventajas de esta técnica son una estimación más robusta del rendimiento del modelo y un uso más eficiente de los datos. Sin embargo, puede ser computacionalmente costosa, especialmente cuando se elige un valor alto para k.

3.2.4. MÉTODO EXPLAINSUMM

En este apartado, presentamos una nueva metodología que aprovecha las tecnologías descritas en el apartado 3.2.3, denominada `ExplainSumm`[11]. El propósito de este método es desarrollar una estrategia que detecte métricas y rasgos predictivos de la calidad de un resumen. Así, tiene como objetivo final identificar características relevantes en la generación de resúmenes automáticos, abordando directamente las preguntas de investigación número 3 (PI3) y 4 (PI4).

Metodología.

1. **Carga y Pre-procesado del Dataset.** En esta etapa, el dataset se carga desde la fuente designada. Una vez cargado, se efectúa el proceso de pre-procesado de datos definido en el párrafo "Pre-procesamiento" de la Sección 3.2.2, que implica tareas de limpieza, transformación y normalización de los datos. Previamente, se ha llevado a cabo también una reducción de memoria[12] del dataset para agilizar todas las futuras operaciones.
2. **Generación de métricas y rasgos automáticos** que permitan evaluar la calidad de un resumen automático desde diversos ángulos. Estas métricas y rasgos vienen definidas en el párrafo "Feature Engineering" de la Sección 3.2.2, y se clasifican en:
   - **Métricas de Evaluación de Resúmenes:** Estas métricas emplean puntuaciones como **ROUGE** y **BLEU** para comparar resúmenes con sus referencias. Adicionalmente, se utilizan embeddings para medir similitudes a través de la **Similitud del Coseno**. La función de evaluación para un resumen (ya sea A o B) y su respectivo resumen de referencia es:
     - $f(\text{Resumen}_A, \text{Resumen}_{\text{Ref.}})$ ROUGE, BLEU, CosSimLLMs
     - $f(\text{Resumen}_B, \text{Resumen}_{\text{Ref.}})$ ROUGE, BLEU, CosSimLLMs
   - **Métricas de Correspondencia Textual:** Estas métricas evalúan la similitud entre el texto original y el resumen, haciendo uso de la **Similitud del Coseno**, el **Ratio de Compresión** y la **Similitud de Jaccard**. La función correspondiente es:
     - $f(\text{Resumen}_A, \text{Text}_{\text{Input}})$ Jaccard Similarity, CosSimLLMs, Compresion Ratio
     - $f(\text{Resumen}_B, \text{Text}_{\text{Input}})$ Jaccard Similarity, CosSimLLMs, Compresion Ratio
   - **Rasgos de Legibilidad y Complejidad Lingüística:** Estos rasgos analizan la estructura del resumen utilizando recuentos lingüísticos y el **Flesch Reading Ease Score (RE)** para determinar la legibilidad. Las métricas empleadas son:
     - $f(\text{Resumen}_A)$ Comprehension Metrics
     - $f(\text{Resumen}_B)$ Comprehension Metrics
3. **Transformación de las métricas y rasgos automáticos.** Definidas las métricas y rasgos automáticos para cada uno de los resúmenes a comparar, es necesario transformar las mismas para interpretar mejor los resultados finales.
   a) En este momento, partimos con la siguiente estructura de datos para modelar la decisión del anotador, con un total aproximadamente de más de 150k observaciones. La columna "Features" hacen referencias a las métricas y rasgos definidas anteriormente, y "Target" a la elección del anotador:

[11] Esta metodología viene definida en clase "ExplainSumm" del siguiente enlace al repositorio GitHub.
[12] Tarea definida en la función "reduceMemory" del siguiente enlace al repositorio GitHub.

Features {Resumen A}	Features {Resumen B}	Target {A, B}

b) Con el objetivo de interpretar los resultados finales, es necesario transformar las "Features" que se acaban de definir. Para ello, se calcula la diferencia entre las respectivas métricas (*spread*) y la media de las mismas (*drift*).

$$\text{Sread} = f(\text{Resumen}_A) - f(\text{Resumen}_B)$$

$$\text{Drift} = \frac{f(\text{Resumen}_A) + f(\text{Resumen}_B)}{2}$$

c) De esta forma, podremos saber, por ejemplo, cuándo una diferencia entre métricas es relevante y a partir de qué niveles puede llegar a ser considerable. Es decir, su interpretabilidad es más directa.

Spread Features {A, B}	Drift Features {A, B}	Target {A, B}

4. **Modelo de clasificación supervisada.** Una vez creadas las características "Spread" y "Drift", que nos permiten interpretar la información de una manera más intuitiva, se define un modelo de clasificación binaria con el objetivo de predecir la decisión del anotador sobre qué resumen representa mejor el texto dado como input.
   - Modelo: El modelo a utilizar es **XGBoostClassifier** con el objetivo de identificar patrones lineales y no lineales. Es considerado un modelo de referencia para datos tabulares. Además, una de las ventajas que ofrece este modelo es la gestión de valores nulos, facilitando todo el proceso.
   - Otro aspecto clave para validar los resultados finales es determinar si el modelo ha sido capaz de reconocer relaciones y patrones (capacidad de generalización) y no simplemente ha memorizado la información (*overfitting*). Para abordar este problema, se utilizan herramientas y técnicas descritas en el apartado 3.2.3, realizando lo siguiente:
     - Separación del dataset en conjuntos de entrenamiento, validación y test.
     - Uso de validación cruzada para la búsqueda de hiperparámetros. En este caso, se utiliza simultáneamente la librería Optuna para la búsqueda de hiperparámetros junto con la validación cruzada.
     - Es esencial asegurarse de que los resultados se mantengan estables a través de las muestras de entrenamiento, validación y test y, finalmente, verificar el poder predictivo en el conjunto de test.
5. **Inteligencia artificial explicable.** En este punto, ya se tiene el modelo de clasificación entrenado y nos apoyamos en él con el objetivo de entender cuáles de las características proporcionadas son las más explicativas y cómo interactúan (dependencia) con la variable a predecir (elección del anotador):
   - Aplicamos la librería de python SHAP (*Shapley Additive exPlanations*): Esto nos permite identificar las métricas o rasgos más relevantes en la toma de decisión del anotador y cómo dichas variables influyen en la predicción del modelo. De esta forma, se comprende cómo el modelo produce dicho output y qué métricas y rasgos son más determinantes.
   - Además, se muestran gráficos de dependencia entre las características y su contribución en la predicción del modelo, lo que posibilita reconocer relaciones entre dichas variables.
6. Tras los pasos anteriores, se han determinado las métricas y rasgos automáticos con mayor poder explicativo de la decisión del anotador (por qué un resumen es mejor que otro), es decir, las características más alineadas con el evaluador.

3.3. EXPERIMENTOS

Los siguientes experimentos se embarcan en la misión de responder a las interrogantes planteadas en las Preguntas de Investigación (Sección 1.2).

El primer experimento (Sección 3.3.1) aborda la PI1 y PI2, explorando la comparación entre los resúmenes automáticos y los resúmenes de referencia generados por humanos. Se busca entender

si las tecnologías actuales han avanzado al punto donde sus resúmenes son indistinguibles, o incluso superiores, a los resúmenes realizados por personas. Además, se cuestiona la relevancia y efectividad de las métricas de evaluación tradicionales, como ROUGE y BLEU, en este escenario contemporáneo.

En el segundo experimento (Sección 3.3.2), se examina el mérito relativo de diferentes métodos de retroalimentación en la evaluación de resúmenes automáticos, ponderando el feedback comparativo frente a las escalas de tipo Likert. Este análisis tiene como objetivo responder a cuestiones subyacentes relacionadas con la percepción humana de la calidad de los resúmenes y cómo evaluarlos de manera más precisa.

El tercer experimento (Sección 3.3.3) responde a la PI3 y PI4, profundizando en la posible influencia de la calidad inherente de los resúmenes en la predictibilidad de ciertas métricas y rasgos. Mediante el uso del método `ExplainSumm` (3.2.4), se busca discernir si ciertos atributos de los resúmenes están correlacionados con su calidad percibida y cómo esta relación varía según el nivel de calidad del resumen.

En conjunto, estos experimentos buscan aportar claridad al campo del resumen automático, proporcionando ideas y orientación en la evaluación y mejora de esta tecnología fundamental. Con una comprensión más profunda de estos aspectos, es posible avanzar hacia sistemas de resumen automático más eficaces y alineados con las expectativas y necesidades humanas.

3.3.1. Experimento 1: Evaluación y Preferencia de Resúmenes Automáticos frente a Resúmenes Humanos

Objetivo. El principal objetivo de este experimento es abordar la pregunta (PI1): ¿Son aquellos resúmenes generados por modelos más complejos (arquitectura / número de parámetros), propios del contexto actual, indistinguibles de resúmenes de referencia o incluso mejores? Además, en el ámbito de evaluación, es esencial cuestionar la relevancia y efectividad de las métricas tradicionalmente utilizadas. Es por ello que también el objetivo de este experimento se centra en responder a esta otra pregunta (PI2): ¿Son válidas las métricas de evaluación tradicionales (ROUGE, BLEU) en el contexto actual, en el cual el evaluador no es capaz de distinguir resúmenes automáticos de resúmenes de referencia (juicio humano)?

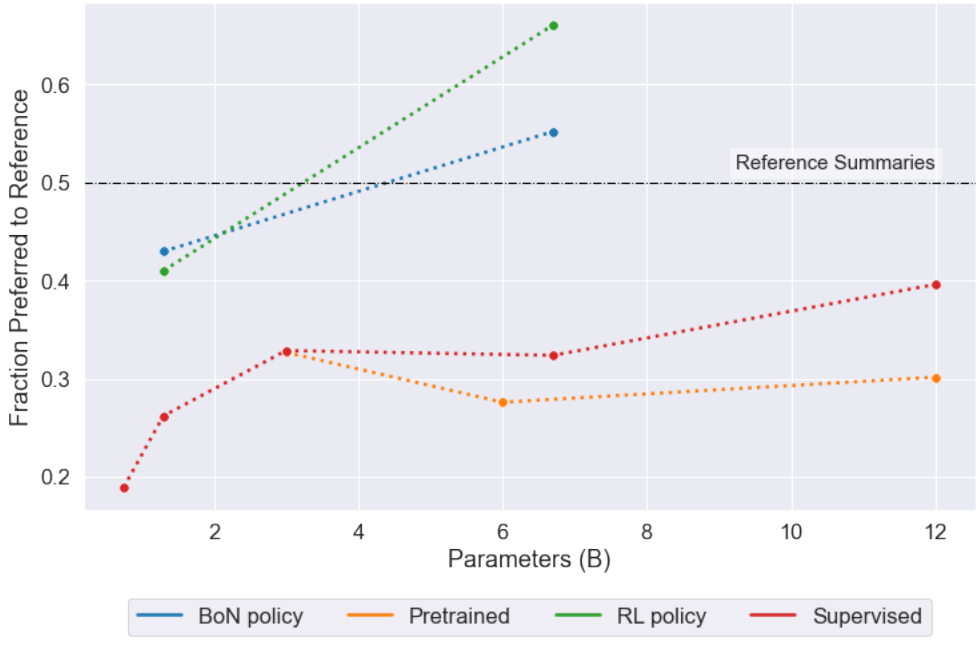

Figura 2: Grado de preferencia sobre los resúmenes de referencia o porcentaje de veces en las que se elije un resumen generado frente a un resumen de referencia (eje x) frente al número de parámetros de cada modelo expresado en *billions* o miles de millones (eje y).

Descripción. Para responder la pregunta planteada (PI1) es esencial identificar para cada una de las observaciones de nuestro dataset el tipo de algoritmo implementado para cada resumen generado. Esto significa que en cada observación, formada por dos resúmenes a comparar, habrá hasta dos algoritmos de generación de resumen diferentes (3.2.1). Una vez identificado el número de parámetros y la tipología del modelo generativo, se filtran aquellas observaciones en las que se compara, un resumen de referencia (generado por un humano) y un resumen generado por un modelo de resumen automático. De esta forma, se puede calcular de forma directa el número de veces que el anotador prefiere un resumen generado por un algoritmo automático frente a un resumen generado por un anotador.

Los resultados se pueden apreciar en la Figura 2, donde se aprecia en términos generales que cuanto mayor es el número de parámetros del modelo automático mayor es su grado de preferencia frente a un resumen generado por un anotador humano. De la misma forma, se muestran 4 tipos diferentes de modelos, donde aquellos más sofisticados (*BoN policy* y *RL policy*), entrenados a través de aprendizaje por refuerzo humanos (*Human feedback models*) obtienen mejores resultados. En el artículo de referencia Stiennon et al. 2020 se llegan a resultados similares.

A la vista de los resultados y teniendo en cuenta que nuestra muestra de observaciones es grande ($>$100k), podemos concluir que a medida que la sofisticación de los modelos de resúmenes automáticos —entendida en términos de número de parámetros y tipología de modelo— aumenta, también lo hace su preferencia sobre resúmenes generados por humanos. De hecho, los resúmenes generados llegan a ser indistinguibles de aquellos creados por humanos. Esta constatación subraya una necesidad apremiante: evolucionar las métricas de evaluación de resúmenes automáticos. Muchas de estas métricas están fundamentadas en, o toman como referencia, resúmenes anotados por humanos.

Dada esta perspectiva, resulta esencial revisar la validez de nuestras métricas actuales. Si los resúmenes automáticos son indistinguibles de los humanos, ¿son nuestras métricas tradicionales (como ROUGE y BLEU) aún adecuadas para evaluar su calidad? Esta reflexión nos lleva a la segunda cuestión de investigación (PI2).

Para responder a esta inquietud, se analiza la correlación entre el grado de preferencia sobre los resúmenes de referencia y las métricas de evaluación, utilizando dos escenarios distintos que permiten explorar la alineación entre estos dos elementos.

- Primer Escenario —Restricción de Muestra: En este escenario, se han excluido las políticas de Aprendizaje por Refuerzo (RL) y BoN. La atención se centra en aquellos casos donde la preferencia hacia los resúmenes automáticos es menor al 40 %. Esta configuración se diseñó para estudiar cómo se comportan las métricas tradicionales cuando los resúmenes generados no son la opción favorita sobre los resúmenes de referencia, lo que podría indicar una menor coherencia con las expectativas humanas.
- Segundo Escenario —Muestra Completa: Aquí, se analiza la muestra completa sin excluir ninguna política o aplicar restricciones en la fracción preferida. Esta vista holística permite evaluar cómo las métricas de evaluación se correlacionan con las preferencias humanas a lo largo de todo el rango de calidad de resúmenes. Se busca una relación monótona positiva entre las métricas y el juicio humano, señalando áreas potenciales donde las métricas actuales pueden no reflejar adecuadamente la calidad percibida por los evaluadores.

El análisis de estos escenarios es esencial para discernir cómo las métricas de evaluación actuales se relacionan con las percepciones humanas. La diferenciación entre ambos escenarios permite tanto una visión detallada como una general, proporcionando un panorama completo de la correlación en cuestión.

La Figura 3 ilustra la correlación entre las preferencias humanas y las métricas de evaluación para ambos escenarios. Las conclusiones que se pueden extraer, son las siguientes:

1. La correlación entre el grado de preferencia humano y las métricas de evaluación son claramente diferentes en las muestras. De esta forma, aquellas métricas consideradas como *gold standard* (ROUGE y BLEU) están perfectamente alineadas con la decisión del anotador cuando se evalúa la muestra restringida (grado de preferencia $<$ 40 %). Estos conjuntos son monótonamente crecientes, a mayor valor de dichas métricas, mayor grado de preferencia sobre los resúmenes de referencia.

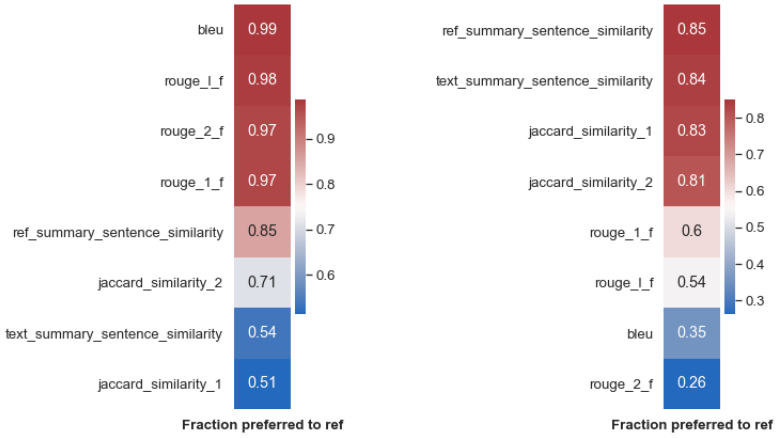

Figura 3: Correlación entre el grado de preferencia sobre los resúmenes de referencia y las métricas de evaluación. A la izquierda: Primer Escenario (Restricción de Muestra) con exclusión de ciertas políticas y foco en resúmenes automáticos menos preferidos. A la derecha: Segundo Escenario (Muestra Completa) que incluye todos los datos sin restricciones. Este análisis busca entender la alineación entre métricas y juicio humano en diferentes contextos.

2. Sin embargo, cuando se expande la muestra, introduciendo aquellas muestras con mayor grado de preferencia humano, los resultados cambian drásticamente. Aquellas métricas que parecían estar perfectamente alineadas (ROUGE y BLEU) con la decisión del anotador resultar ser las menos efectivas al considerar toda la muestra.
3. Las métricas más efectivas, entendiéndose como efectiva la correlación positiva entre el grado de preferencia humano y las métricas de evaluación, son aquellas métricas que utilizan la similitud del coseno entre los embeddings del texto/resumen de referencia y el resumen dado. Así mismo las métricas de *Jaccard*, son también de las métricas más efectivas siendo aparentemente de las que menos en la muestra reducida. Queda claro, que la métrica mas robusta a lo largo de ambas muestras es la métrica de similitud coseno entre los embeddings del texto y el resumen dado.

Los resultados presentados en las Figura 3 son datos a nivel agregado, pero con el objetivo de entender mejor los resultados, se muestra a continuación en la Figuras 4 y 5 la evolución de las métricas a lo largo de los diferentes grados de preferencia sobre los resúmenes de referencia (mismos datos pero con otra perspectiva).

La métrica BLEU que aparece como la métrica más alineada con la decisión de los anotadores en el primer escenario planteado deja de serlo al considerar toda muestra. Este fenómeno se aprecia perfectamente en la parte superior izquierda de la Figura 4. A su vez, tal y como se muestra en la Figura 2, si solo se consideraran resúmenes de baja calidad (grado de preferencia sobre los resúmenes $<$ 40 %) el grado de alineamiento sería del 99 %.

Otro grupo de métricas consideras como métricas de referencia, es ROUGE. Como en el caso anterior, esta familia de métricas parece estar también perfectamente alineadas con los anotadores. Sin embargo, cuando se considera la muestra completa, el grado de alineamiento cae drásticamente. Este fuerte decaimiento se aprecia claramente en los gráficos de la Figura 4. Aquellas métricas que parecían ser relevantes a la hora evaluar resúmenes, dejan de serlo cuando la calidad de los mismos mejora sustancialmente, dejando de estar alineadas con el anotador.

Al observar las demás métricas presentadas, algunas que inicialmente no parecían ser las más adecuadas en términos de alineamiento resultan ser las más idóneas al considerar el conjunto completo de la muestra. Se dispone de métricas de diversa sofisticación. Por ejemplo, la métrica de *Jaccard* muestra una correlación con la decisión del anotador superior al 80 %. Sin embargo, las métricas que miden la similitud del coseno entre los diferentes embeddings generados, tanto para textos como para resúmenes, resultan ser las más satisfactorias. En este contexto, la métrica que evalúa la similitud textual mediante embeddings pre-entrenados entre el texto y el resumen destaca por su excelente

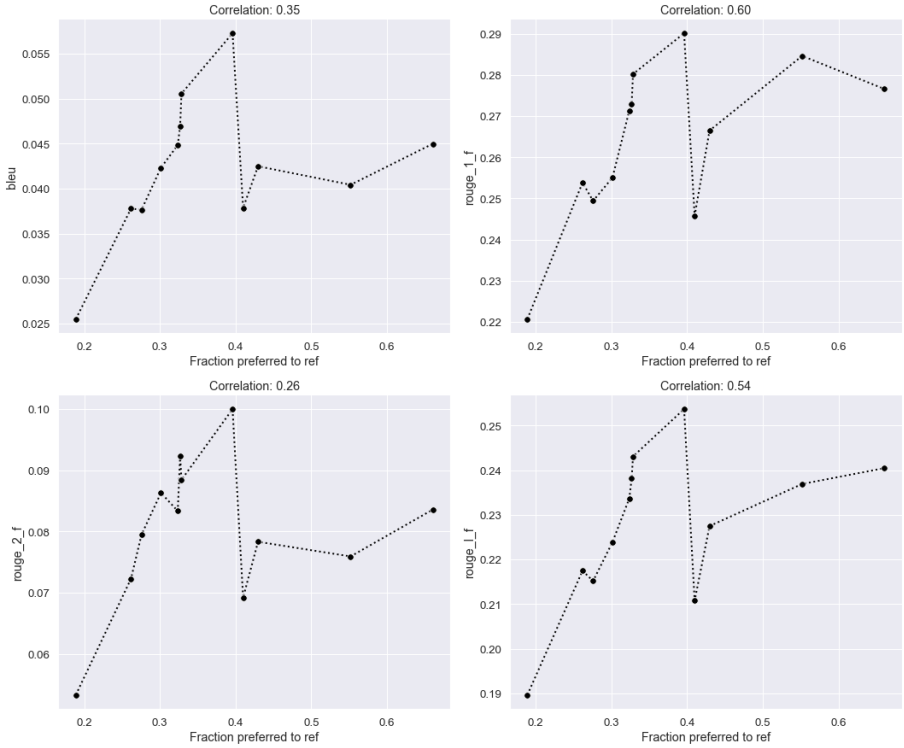

Figura 4: Gráfico de líneas entre el grado de preferencia sobre los resúmenes de referencia y las diferentes métricas de evaluación BLEU y ROUGE.

comportamiento en ambas muestras. Este hecho se evidencia en el gráfico 5, donde se observa una clara relación monótona positiva entre el grado de preferencia de los resúmenes respecto a los de referencia y el valor de la métrica de similitud.

3.3.2. Experimento 2: Análisis de Métricas de Evaluación de Resúmenes Automáticos en Diferentes Colecciones de Datos Anotados

Objetivo. Investigar cómo las métricas de evaluación de resúmenes automáticos se correlacionan con diferentes tipos de feedback humano: feedback comparativo y escalas psicométricas tipo Likert.

Descripción. Este análisis busca sustentar la idea de que el feedback comparativo (comparaciones directas entre resúmenes) puede proporcionar percepciones más valiosos que el feedback obtenido a través de escalas psicométricas tipo Likert. Varios estudios han respaldado la superioridad del feedback comparativo en la optimización de LLMs (Li et al. 2019; Stiennon et al. 2020; Touvron et al. 2023).

Para este experimento:

- Mediremos la correlación entre la escala Likert y el valor de las métricas de evaluación.
- Usando los datos del Experimento 1 (3.3.1), evaluaremos la correlación entre el Grado de Preferencia sobre la Referencia y el valor de las métricas de evaluación tal como se presenta en la Figura 2.

Los datos sugieren que las métricas basadas en "similitud de coseno a través de embeddings" están en mayor consonancia con las evaluaciones humanas. No obstante, se observa discordancia entre otras métricas. Ambos enfoques de feedback, comparativo y *Likert*, podrían ser válidos dado que ambos provienen de evaluadores humanos. Sin embargo, usar un punto de referencia (comparaciones de resúmenes) facilita la decisión de qué resumen es superior. Esta variabilidad en las evaluaciones se evidencia en la dispersión de las anotaciones en la escala *Likert*, como se muestra en la Figura 7.

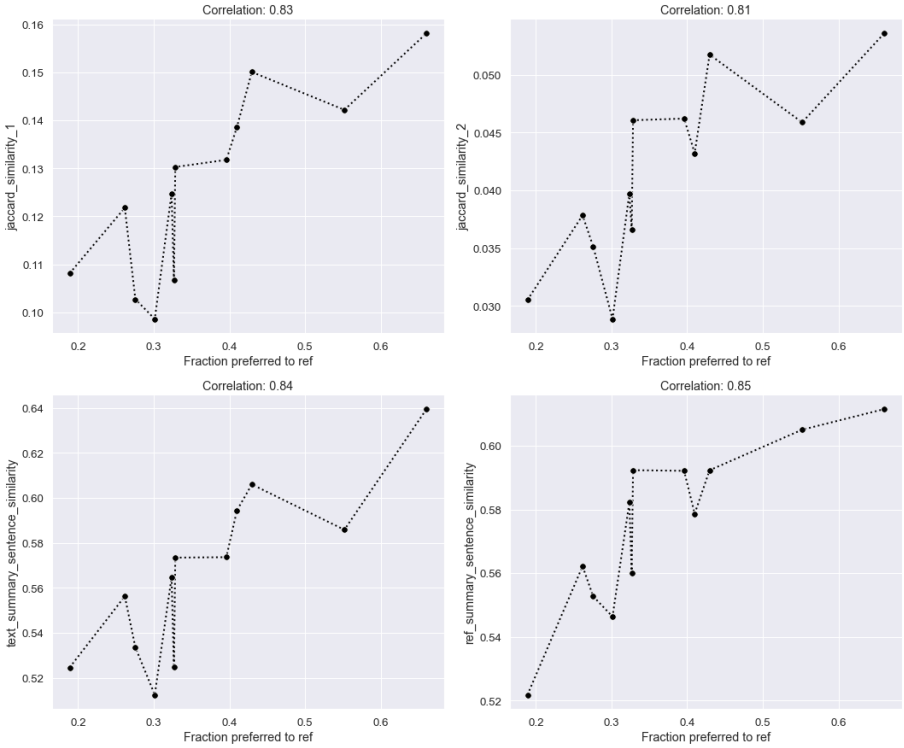

Figura 5: Gráfico de líneas entre el grado de preferencia sobre los resúmenes de referencia y las diferentes métricas de evaluación de Jaccard y la similitud del coseno entre los embeddings texto-resumen y resumen-resumen.

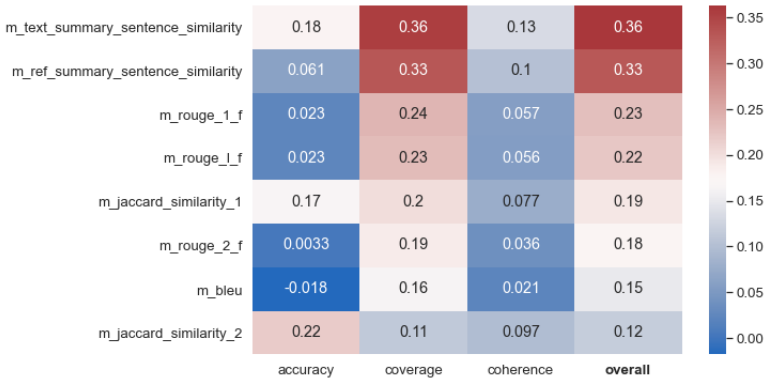

Figura 6: Correlación entre la escala *Likert* y el valor de las métricas de evaluación.

Basándonos en lo anterior y en el Experimento 1 (3.3.1), junto con la literatura citada, concluimos que la meta-evaluación de métricas usando feedback comparativo puede ser más esclarecedora al determinar qué métricas de evaluación de resúmenes automáticos se alinean mejor con la percepción humana.

3.3.3. Experimento 3: Evaluación de la Influencia de la Calidad del Resumen en la Capacidad Predictiva de Métricas y Rasgos

Objetivo. El propósito principal de este tercer y último experimento es abordar dos preguntas fundamentales:

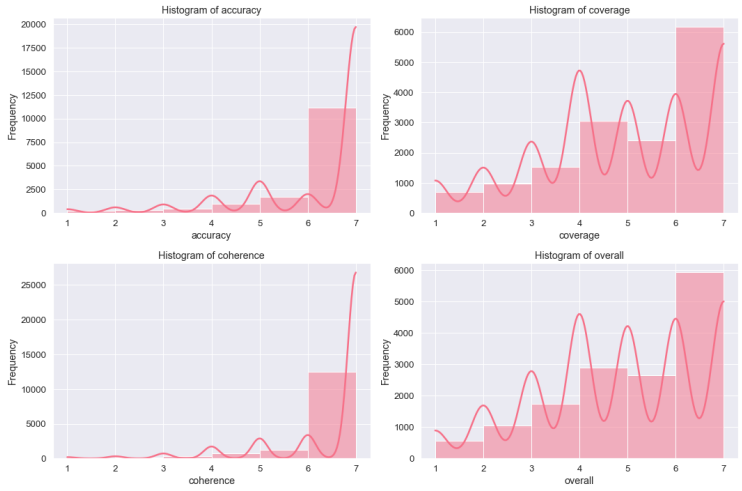

Figura 7: Histograma de las escalas tipo *Likert* en el dataset anotado.

- **PI3**: ¿Es la propia calidad de los resúmenes a comparar una causa directa de la perdida de capacidad predictiva de ciertos rasgos y/o métricas?
- **PI4**: ¿Qué métricas o rasgos mantienen su capacidad predictiva conforme se escala la calidad de los resúmenes?

Para responder a estas cuestiones, se utiliza el método `ExplainSumm`.

Descripción. Para determinar si la propia calidad de los resúmenes a comparar es una causa directa de la perdida de capacidad predictiva de ciertas métricas y rasgos, se segmenta el dataset atendiendo al grado de preferencia de los resúmenes generados por distintos modelos en relación con los resúmenes de referencia. Como se muestra en la Figura 2, cada modelo de resumen automático, en función de su tipología y número de parámetros, presenta diferentes grados de preferencia, reflejando así distintas calidades de resúmenes, llegando incluso a niveles comparables con los generados por anotadores humanos (grado de preferencia sobre la referencia $\geq 50\,\%$).

Tras identificar los niveles de calidad de cada modelo, segmentamos el dataset basándonos en el origen de los resúmenes a comparar en cada observación, aislando así las observaciones donde se comparan resúmenes de modelos con calidades similares. Segmentamos el dataset *comparisons* en tres grupos:

1. Resúmenes de alta calidad *(high)*
2. Resúmenes de calidad media *(mid)*
3. Resúmenes de baja calidad *(low)*

Con el dataset ya segmentado, se aplica el método `ExplainSumm` descrito en las Sección 3.2.4 a cada grupo. Para evaluar la robustez de los resultados obtenidos con este método, comprobamos los niveles de precisión (accuracy) en cada uno de los conjuntos de *test* definidos para cada muestra (representando el 25 % del total de cada muestra).

Cuadro 1: Métricas de evaluación para cada una de las 3 muestras.

Metric	Low Quality	Mid Quality	High Quality
Accuracy	62.69 %	60.52 %	59.33 %
Precision	62.09 %	61.74 %	63.42 %
Recall	62.74 %	61.25 %	59.89 %
F1 Score	62.41 %	61.49 %	61.6 %
ROC AUC	62.69 %	60.5 %	59.28 %

La Tabla 1 evidencia una estabilidad en términos de *accuracy* a lo largo de las tres muestras, si bien se percibe una ligera disminución del poder predictivo a medida que aumenta la calidad de los resúmenes. A continuación, se calculan los valores SHAP para entender más profundamente cómo cada característica contribuye a la predicción del modelo.

Al analizar los resultados presentados en la Figura 8, es posible realizar las siguientes observaciones:

- Las métricas de similitud textual basadas en los embeddings de modelos pre-entrenados, junto con las métricas de similitud de Jaccard, emergen como las características más indicativas de la calidad de un resumen. Estas métricas sugieren que los resúmenes con valores más elevados son más propensos a ser seleccionados como el resumen ideal por los anotadores.
- Este patrón es especialmente evidente en las muestras de baja y media calidad, donde ciertas métricas, como el "spread" de Similitud de Oraciones entre Texto y Resumen, Similitud de Oraciones entre Resumen de Referencia y Resumen, y Similitud de Jaccard para N-gramas de Tamaño 1 destacan como las más predictivas. No obstante, en las muestras de alta calidad, la clara influencia de estas métricas se diluye, pese a que el "spread" de Similitud de Jaccard para N-gramas de Tamaño 1 y Similitud de Oraciones entre Texto y Resumen continúan demostrando una fuerte robustez a lo largo de las tres muestras, como se puede corroborar al analizar la media absoluta de los valores SHAP presentes en la Figura 9.
- Es notable que, a medida que se analizan modelos de resúmenes automáticos más avanzados, los rasgos específicos adquieren mayor relevancia en comparación con las métricas en general. Por ejemplo, la cantidad de monosílabos, el número de palabras en el texto y la relación de compresión entre texto y resumen se alinean más estrechamente con las preferencias del anotador. Este hallazgo sugiere que, a medida que se incrementa la complejidad, la elección del anotador se vuelve menos predecible y se ve influenciada por una combinación más diversa de factores.
- Con respecto a las métricas clásicas en la literatura, como ROUGE y BLEU, se observa que no son predictivas inicialmente, ganando cierta relevancia sólo cuando se evalúan resúmenes de baja calidad.
- Desde la perspectiva del autor, es destacable el robusto poder predictivo de la métrica "spread" de Similitud de Jaccard para N-gramas de Tamaño 1. A pesar de ser una métrica tradicional en la evaluación de resúmenes automáticos, se mantiene como una de las más resilientes, comparable con métricas derivadas de modelos pre-entrenados significativamente más sofisticados, como el "spread" de Similitud de Oraciones entre Texto y Resumen y Similitud de Oraciones entre Resumen de Referencia y Resumen.

Por último, con el propósito de analizar más a fondo cómo estas métricas y rasgos influyen en la decisión del anotador, se visualizan los gráficos de dependencia SHAP para cada característica del conjunto de datos en las Figuras 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 20, y 21. Estos gráficos ofrecen una perspectiva más clara sobre la magnitud y el tipo de dependencia de cada métrica y rasgo con respecto a la decisión del anotador. En este contexto, es relevante destacar lo siguiente:

- Sin duda, tal como se confirma en la Figura 9 donde las características están ordenadas por importancia descendente, la métrica "spread" de Similitud de Oraciones entre Texto y Resumen posee el mayor poder predictivo. Esto se evidencia en el gráfico de dependencia (Figura 10), donde se observa una relación monótona entre el valor de la métrica y la decisión del anotador. Esta relación se mantiene estable a lo largo de las tres muestras. Por otro lado, la métrica "spread" de Similitud de Oraciones entre Resumen de Referencia y Resumen es predictiva en las muestras de calidad de resúmenes *low* y *mid*. Sin embargo, pierde relevancia al comparar resúmenes de alta calidad, como se observa en la figura 11. Es coherente pensar que esta métrica reduzca su poder predictivo a medida que los resúmenes generados se asemejan o son indistinguibles de los resúmenes de referencia.
- El siguiente conjunto de variables con más influencia a lo largo de las tres muestras son las métricas de similitud de Jaccard. Estas métricas evalúan la intersección sobre la unión de conjuntos, en este caso, texto y resumen proporcionado. Es notable la importancia que tiene esta métrica en la explicación de la decisión del anotador. Sin embargo, en la muestra de resúmenes de alta calidad, no se observa tan claramente la relación que tiene en los otros conjuntos, como se muestra en los gráficos 12 y 13.

- Respecto a las métricas consideradas como *gold standards*, ROUGE y BLEU, se muestra en los gráficos de dependencia 14, 15, 16, y 17 como su poder explicativo es limitado en este estudio, y como existe un fuerte decaimiento en su poder predictivo cuando elevamos la calidad de los resúmenes en la meta-evaluación. Este efecto es común a todas las métricas de la familia ROUGE y BLEU.
- Por último, profundizando en los rasgos, se aprecia en los gráficos de dependencia 18, 19, 20, y 21 que ninguno de los rasgos son estables (mismo tipo de dependencia) a lo largo de las 3 muestras. Y no solo no son estables, sino que además tienen relaciones completamente contrarias a medida el grado de calidad de los resúmenes cambia. Resulta curioso, como ya se ha comentado anteriormente, que en métricas que miden el número de palabras o monosílabos en el resumen, "tener más" implica un mayor alineamiento con el anotador en el dataset *high*, y lo contrario en la muestra *low*.

3.4. Discusión general

Dentro del ámbito de la generación y evaluación de resúmenes automáticos, se han llevado a cabo tres experimentos significativos para entender mejor el panorama actual y el alcance de las métricas utilizadas. A continuación, presentamos una revisión concisa de los objetivos y hallazgos más relevantes de cada experimento.

3.4.1. Experimento 1: Evaluación y Preferencia de Resúmenes Automáticos frente a Resúmenes Humanos

En este experimento se busca comparar la calidad y efectividad de los resúmenes producidos por modelos avanzados de IA con aquellos creados manualmente por expertos. Específicamente, se pretende: (1) Explorar si los resúmenes generados por modelos de resumen automático son comparables o incluso superiores a los resúmenes creados por expertos humanos, y (2) Investigar la idoneidad de métricas de evaluación tradicionales (como ROUGE y BLEU) en el contexto actual, donde puede ser complicado diferenciar entre un resumen automático y uno humano.

Descubrimientos Clave. A partir de la experimentación y el análisis de los datos recogidos, se destacan las siguientes observaciones clave:

- Se observa que a medida que aumenta la complejidad y el número de parámetros de los modelos automáticos de resumen, mayor es su grado de preferencia de elección frente a un resumen generado por un anotador humano (Figura 2).
- En situaciones donde los modelos de generación automática de resúmenes son especialmente avanzados, la distinción entre sus outputs y los resúmenes humanos puede volverse ambigua, e incluso indistinguible.
- Métricas tradicionalmente aceptadas y empleadas para evaluar resúmenes automáticos, como ROUGE y BLEU, muestran menor correlación con la percepción humana en resúmenes de alta calidad.
- La métrica que utiliza la similitud coseno entre los embeddings del texto y el resumen utilizando la librería `SentenceTransformers`, destaca por ser la que mejor se alinea con las evaluaciones humanas en una variedad de escenarios, sugiriendo su potencial superioridad para evaluar resúmenes automáticos.

Los resultados del Experimento 1 subrayan la creciente capacidad de los modelos de IA avanzados en la generación de resúmenes, alcanzando e incluso superando en ciertos contextos la calidad de los resúmenes humanos. No obstante, se pone de manifiesto la necesidad de reconsiderar y adaptar las métricas tradicionales de evaluación, dada la disminución en su correlación con la percepción humana en escenarios de alta calidad. La métrica basada en la similitud coseno junto con `SentenceTransformers` emerge como una herramienta prometedora para la evaluación contemporánea de resúmenes automáticos, marcando un posible camino a seguir en futuras investigaciones en este ámbito.

3.4.2. Experimento 2: Análisis de Métricas de Evaluación de Resúmenes Automáticos en Diferentes Colecciones de Datos Anotados

El propósito principal de este segundo experimento es investigar la correlación entre métricas de evaluación y distintos tipos de feedback humano. Se pone especial énfasis en el feedback comparativo y las escalas tipo Likert.

Descubrimientos Clave. A través de la aplicación de métodos analíticos presentados en la Sección 3.3.2, los descubrimientos clave en esta etapa del estudio son los siguientes:

- El feedback comparativo, donde se realizan comparaciones directas entre resúmenes, ha demostrado ser más informativo y revelador que las evaluaciones obtenidas mediante las escalas tipo Likert.
- Las métricas que se basan en la similitud de coseno utilizando embeddings parecen ser las más alineadas con las percepciones humanas. Estas métricas, al emplear vectores semánticos que capturan el significado de las palabras y las frases en los resúmenes, ofrecen una aproximación más fidedigna a cómo los humanos interpretan y evalúan el contenido.
- Existe una discordancia notoria entre las métricas tradicionales y cómo estas se comparan con las evaluaciones humanas. Esto sugiere la necesidad de revisar y posiblemente refinar nuestras métricas actuales.
- Aunque ambos tipos de feedback, comparativo y Likert, son considerados válidos porque reflejan las opiniones de evaluadores humanos, el feedback comparativo ofrece una claridad que facilita la identificación de resúmenes superiores. Es más directo y elimina cierto grado de ambigüedad que puede surgir al usar escalas.

En conclusión, el Experimento 2 ha proporcionado ideas valiosas sobre la eficacia y relevancia de diferentes métricas de evaluación en el contexto de la generación de resúmenes automáticos. Los hallazgos subrayan la importancia del feedback comparativo como herramienta de evaluación y señalan áreas de mejora en nuestras métricas de evaluación actuales. Estas percepciones, al lado de la evidente alineación de ciertas métricas con la percepción humana, sirven como base para futuras investigaciones y refinamientos en el campo de evaluación de resúmenes. Estas lecciones aprendidas permitirán ajustes metodológicos y prácticos que mejorarán la precisión y fiabilidad de nuestras evaluaciones en futuros experimentos.

3.4.3. Experimento 3: Evaluación de la Influencia de la Calidad del Resumen en la Capacidad Predictiva de Métricas y Rasgos

El tercer experimento se embarca en la tarea de entender cómo la variabilidad en la calidad de los resúmenes puede influir en las métricas y rasgos empleados para evaluarlos. En concreto, se busca: (1) Investigar la relación entre la calidad de los resúmenes y la capacidad predictiva de las métricas y rasgos empleados, y (2) determinar qué métricas o rasgos mantienen su relevancia y poder predictivo a través de diferentes calidades de resúmenes.

Descubrimientos Clave. Basados en los resultados obtenidos de la aplicación del método `ExplainSumm` y del análisis exhaustivo de las diferentes calidades de resúmenes (alta, media y baja), los hallazgos clave incluyen:

- Una estabilidad notable en la precisión a través de las categorías de calidad. Sin embargo, es importante resaltar que la capacidad predictiva tiende a disminuir para resúmenes de alta calidad.
- Las métricas que examinan la similitud textual basadas en embeddings de modelos pre-entrenados y la similitud de Jaccard se erigen como indicativos fundamentales para determinar la calidad de un resumen. Estas métricas reflejan que resúmenes con valores elevados son más propensos a ser seleccionados como ideales por los anotadores.
- Las métricas ROUGE y BLEU, a menudo citadas como métricas estándar en la literatura de generación de resúmenes, no demostraron ser tan predictivas inicialmente. Su relevancia aumenta solamente cuando se analizan resúmenes de menor calidad.
- Una observación interesante es la robustez del "Spread" de Similitud de Jaccard, que se destaca incluso frente a métricas derivadas de modelos más avanzados y pre-entrenados.

- Se detectó que rasgos específicos, como el conteo de palabras o monosílabos en un resumen, pueden variar en relevancia según la calidad del resumen. En resúmenes de alta calidad, una mayor cantidad de palabras o monosílabos se correlaciona positivamente con la alineación con las preferencias del anotador, mientras que ocurre lo contrario en resúmenes de baja calidad.

En resumen, el Experimento 3 ha arrojado luz sobre la relación intrínseca entre la calidad de los resúmenes y la capacidad predictiva de las métricas y rasgos utilizados para evaluarlos. Los hallazgos subrayan la importancia de considerar la variabilidad en la calidad de los resúmenes al seleccionar y aplicar métricas y rasgos de evaluación. Es evidente que, a medida que avanzamos en la generación y evaluación de resúmenes, es esencial una comprensión matizada de cómo la calidad influye en la evaluación, asegurando así que los métodos y herramientas de evaluación sean tanto precisos como relevantes para sus propósitos.

4. Resultados y discusión

Este capítulo se dedica a desentrañar los hallazgos y resultados emergentes de nuestra investigación, contrastando su significado y relevancia en el contexto más amplio del campo de resúmenes automáticos. El objetivo principal es situar nuestros hallazgos en el panorama general de investigaciones anteriores, identificando tanto congruencias como desviaciones notables, así como discutir las implicaciones significativas que surgen de nuestros datos, tanto a nivel práctico como teórico.

En la Sección 4.1, establecemos un diálogo con estudios previos, identificando convergencias y divergencias en nuestros hallazgos. Esta comparación nos permite no solo contextualizar nuestros resultados, sino también entender nuestra contribución única al cuerpo de conocimiento existente.

Posteriormente, en 4.2, profundizamos en las ramificaciones de nuestros hallazgos. Desgranamos las implicaciones prácticas, proporcionando ideas útiles para la implementación y mejora de sistemas de resumen automático. Además, exploramos las consecuencias teóricas, delineando cómo nuestros resultados pueden influir en la concepción teórica del dominio.

Por último, en 4.3, adoptamos una postura autocrítica, evaluando las limitaciones de nuestro enfoque y diseño experimental. Reconocer y abordar estas limitaciones no solo refuerza la integridad de nuestra investigación, sino que también sienta las bases para futuras investigaciones en el área.

Con todo, esta sección se esfuerza por presentar una visión holística y balanceada de nuestros hallazgos, situándolos adecuadamente en el panorama actual de la generación y evaluación de resúmenes automáticos.

4.1. Comparación con Estudios Previos

Para una evaluación efectiva y constante mejora de los modelos de generación automática de resúmenes, es esencial tanto analizar en profundidad los resultados actuales como contrastarlos con investigaciones anteriores. Esta sección tiene como objetivo comparar los resultados de nuestros experimentos con los estudios y tendencias anteriores presentes en la literatura. De esta manera, buscamos situar nuestros resultados dentro del cuerpo de conocimientos existentes, identificar alineaciones, desviaciones y, lo más importante, discernir las contribuciones únicas de nuestra investigación.

Experimento 1: Evaluación y Preferencia de Resúmenes Automáticos frente a Resúmenes Humanos. A partir de los resultados del **Experimento 1**, es evidente que (1) la evaluación de resúmenes automáticos debe transicionar en una dirección diferente a la comparación entre resúmenes generados por modelos y resúmenes humanos, y (2) las métricas de evaluación tradicionales como ROUGE y BLEU no son efectivas en el contexto actual. A continuación se comparan estos resultados con las investigaciones anteriores en el área:

- **Preferencia de Resúmenes Automáticos frente a Resúmenes Humanos**: Nuestros hallazgos indican que los resúmenes generados por modelos avanzados son en muchos casos indistinguibles o incluso preferidos sobre los resúmenes humanos. Esto respalda el estudio Stiennon et al. 2020, que sugiere que los resúmenes basados en feedback humano superan en desempeño a los resúmenes de referencia humanos. No obstante, esto contrasta con la

perspectiva tradicional en la que el juicio humano es considerado el estándar de oro para la evaluación, tal como se menciona en la Sección 2.3.

- **Efectividad de métricas tradicionales**: Hemos cuestionado la relevancia de métricas como ROUGE y BLEU en el contexto actual. Esta postura se alinea con la discusión en la Sección 2.3, donde estudios como Novikova et al. 2017 y Schluter 2017 también cuestionan la correlación de estas métricas con el juicio humano, especialmente en sistemas de generación automática de lenguaje mediante modelos avanzados.
- **Similitud del coseno**: La métrica de similitud del coseno, que compara los embeddings del texto y el resumen usando la librería `SentenceTransformer` (identificada como la más robusta y correlacionada con la percepción humana en nuestro experimento), no ha sido abordada en detalle en las investigaciones previas mencionadas. Esto indica que podría ser una dirección interesante para exploraciones futuras, brindando una perspectiva fresca en la evaluación de resúmenes automáticos.

En resumen, a pesar del consenso emergente sobre la limitada robustez de las métricas tradicionales en el panorama actual, la introducción de nuevas métricas y la comparativa con resúmenes humanos subraya la imperativa necesidad de reevaluar y adaptar continuamente nuestras aproximaciones en el ámbito de la generación automática de resúmenes.

Por otra parte, a nuestro juicio, y debido a la rápida evolución de la calidad de los resúmenes generados de manera automática, se debería considerar dejar de lado los resúmenes de referencia para la evaluación y transicionar hacia métricas libres de referencia. En nuestro estudio demostramos que la métrica que compara la similitud semántica entre el texto y el resumen ofrece un mayor grado de alineación con las evaluaciones de anotadores humanos.

Experimento 2: Análisis de Métricas de Evaluación de Resúmenes Automáticos en Diferentes Colecciones de Datos Anotados. Los resultados obtenidos en nuestro **Experimento 2** han arrojado observaciones valiosas en relación con el feedback humano y la correlación entre las métricas automáticas y la percepción humana. A continuación, presentamos una comparación de estos hallazgos con las investigaciones previas en el área:

- En relación con el primer hallazgo, la importancia del feedback humano está en sintonía con la creciente atención que la comunidad ha dado al feedback humano como piedra angular (Böhm et al. 2019; Leike et al. 2018; Li et al. 2019; Stiennon et al. 2020; Ziegler et al. 2019).
- Las observaciones relativas a la correlación de las métricas con las evaluaciones humanas son consistentes con las preocupaciones presentadas en la Sección 2.3. Específicamente, nuestro análisis refuerza los resultados de Novikova et al. 2017 y Fabbri et al. 2020 que indican una correlación imperfecta entre las métricas automáticas y el juicio humano.
- Aunque nuestros resultados sugieren que las métricas basadas en "similitud de coseno a través de embeddings" están más alineadas con las evaluaciones humanas, es importante notar que, tal como mencionan Novikova et al. 2017 y Schluter 2017, no existe una métrica única que sea ideal para todas las situaciones.
- La exploración de las diferentes colecciones de feedback humano presentadas en este trabajo, específicamente (1) Comparación de resúmenes (comparisons) y (2) evaluación de la calidad de los resúmenes (axis) detalladas en la Sección 3.2.1, sugiere que las anotaciones que cuentan con un punto de referencia, como en el caso de las comparaciones, presentan una mayor variabilidad en contraste con las anotaciones basadas en escalas Likert (axis). Esta mayor variabilidad enriquece la inferencia que puede realizarse a partir de los datos. La preferencia por conjuntos de datos anotados con un enfoque comparativo en lugar de anotaciones tipo Likert también encuentra respaldo en otros estudios, como se muestra en Li et al. 2019; Stiennon et al. 2020.

En resumen, si bien nuestros hallazgos son específicos para el dominio y contexto de nuestro estudio, coinciden y enriquecen las investigaciones actuales en la generación automática de resúmenes. En conjunto con la literatura existente, nuestros resultados resaltan la imperativa necesidad de seguir explorando y mejorando las métricas de evaluación y la integración del feedback humano, con el fin de desarrollar sistemas de generación de resúmenes que estén más alineados con el juicio humano.

Experimento 3: Evaluación de la Influencia de la Calidad del Resumen en la Capacidad Predictiva de Métricas y Rasgos. Los resultados del **Experimento 3** que responden a las preguntas 3 y 4 planteadas en la Sección 1.2 presentan hallazgos relevantes desde un punto de vista teórico y práctico. A continuación, se contrastan los resultados clave de nuestro experimento con investigaciones previas en el área:

- La observación de que métricas como ROUGE y BLEU demostraron tener un poder predictivo limitado, sobre todo cuando se evaluaron resúmenes de baja calidad, es coherente con la discusión de Novikova et al. 2017, donde se resalta la correlación débil de estas métricas con las valoraciones humanas. Además, esto resuena con las críticas hacia ROUGE presentadas por Schluter 2017 y Paulus et al. 2017, que señalan posibles fallos en la métrica que podrían favorecer resúmenes menos legibles.
- Las métricas basadas en la similitud del coseno a través de embeddings pre-entrenados, y similitud de Jaccard, fueron identificadas como características clave de la calidad de un resumen en nuestro estudio. Aunque la literatura revisada en la Sección 2.3 se centra principalmente en métricas tradicionales, la influencia y el valor de estas métricas modernas sugieren una dirección para futuras investigaciones en la comunidad.
- La observación de que, a medida que se analizaron modelos de resumen más avanzados, rasgos específicos ganaban relevancia en comparación con métricas generales, resuena con los hallazgos de Chaganty et al. 2018 y Kryscinski et al. 2019. Ambos trabajos enfatizan la necesidad de una evaluación más matizada y holística de los modelos de generación de resumen.
- El descubrimiento de que el "spread" de Similitud de Jaccard para N-gramas de Tamaño 1 es especialmente resiliente, destaca la importancia de métricas basadas en similitud, una dirección que merece más investigación y comparación con métricas derivadas de modelos pre-entrenados, como se sugiere en la Sección 2.3.
- Finalmente, la observación de que algunos rasgos, como el número de palabras o monosílabos en el resumen, mostraron patrones opuestos en categorías de alta y baja calidad, es especialmente intrigante. Esto resalta la complejidad de las características inherentes de los resúmenes y sugiere que no solo las métricas, sino también los rasgos inherentes de los textos, deben ser considerados en la evaluación de la calidad del resumen.

En resumen, los resultados de nuestro **Experimento 3** refuerzan y complementan muchas de las discusiones y hallazgos anteriores en el área, sugiriendo al mismo tiempo nuevas direcciones y enfoques para futuras investigaciones.

Además, `ExplainSumm`, el método desarrollado en este trabajo y aplicado en el presente experimento, demuestra su capacidad para identificar las métricas y rasgos más relevantes a considerar en la generación automática de resúmenes. Esto es válido en cualquiera de las 3 muestras utilizadas, que representan diferentes escenarios de capacidades generativas de resúmenes, desde la menor hasta la mayor calidad. `ExplainSumm` ha probado ser eficiente, especialmente al compararlo con otros experimentos (como los dos primeros). En ellos, otros enfoques, como medir la correlación con el juicio humano, resultan menos efectivos debido a sus limitaciones para detectar patrones no lineales.

4.2. Implicaciones Prácticas y Teóricas

Los resultados obtenidos a lo largo de nuestros experimentos tienen un profundo impacto tanto en la aplicación práctica como en la comprensión teórica del campo de la generación automática de resúmenes. A continuación, se destilan estas implicaciones para orientar futuras investigaciones y aplicaciones.

Implicaciones Prácticas:

- Los sistemas de generación automática de resúmenes han avanzado a un punto donde, en muchos casos, sus resúmenes son indistinguibles o incluso preferidos sobre los resúmenes humanos.

- Las métricas tradicionales como ROUGE y BLEU podrían no ser completamente adecuadas para evaluar estos sistemas avanzados en el contexto actual.
- Se sugiere la similitud coseno entre embeddings y la similitud de Jaccard, basados en embeddings de modelos pre-entrenados, como métricas potencialmente valiosas y robustas para la evaluación práctica de sistemas automáticos. Estas métricas deberían recibir más atención y estudio en la práctica.
- Es esencial integrar el feedback humano no solo en el entrenamiento de modelos de lenguaje, sino también en la meta-evaluación de métricas.
- Se debe prestar atención a la complejidad inherente de los resúmenes (calidad) y considerar rasgos específicos en las evaluaciones.
- Dada la capacidad del método `ExplainSumm` para identificar patrones tanto lineales como no lineales, se resalta la importancia de alejarse de medidas lineales tradicionales, como la correlación entre métricas y juicio humano. Es esencial transicionar hacia modelos más avanzados que puedan capturar de manera efectiva la complejidad inherente en la evaluación de resúmenes.

Implicaciones Teóricas:

- Nuestros hallazgos refuerzan la necesidad de ir más allá de las métricas tradicionales al evaluar sistemas de resumen automático, sugiriendo una revisión y adaptación hacia métricas más novedosas y modernas.
- Sugerimos evidencias empíricas sobre la eficacia relativa de los modelos avanzados en la generación de resúmenes en comparación con los resúmenes humanos, lo que enriquece el cuerpo de conocimiento en este campo.
- Se sugieren nuevas direcciones y enfoques para la investigación futura en el dominio de la generación de resúmenes automáticos, en especial la importancia del feedback humano y la adaptación constante en la evaluación.
- Nuestros estudios complementan y extienden investigaciones previas, ofreciendo una comprensión más matizada y holística de la evaluación en el dominio.

4.3. Limitaciones y Estrategias de Mitigación

En el proceso de evaluación de métricas para resúmenes automáticos, se identifican diversas limitaciones que pueden influir en la validez y aplicabilidad de los hallazgos presentados. Es crucial ser consciente de estas limitaciones para interpretar correctamente los resultados y planificar futuras investigaciones:

- **Sesgo intrínseco en la generación de resúmenes:** Los resúmenes generados pueden compartir patrones y errores similares debido al proceso de entrenamiento. Esta coincidencia puede disminuir la eficacia de las comparaciones y dificultar la identificación de diferencias sutiles entre los resúmenes.
- **Dependencia de modelo:** Los resúmenes originados por modelos con arquitecturas parecidas pueden presentar características similares, lo que podría introducir sesgos en las evaluaciones y limitar la extrapolación de los resultados.
- **Variabilidad subjetiva entre anotadores:** Dado que la calidad de un resumen puede ser interpretada de forma diferente por distintos evaluadores, es posible que existan variaciones en las evaluaciones, lo que añade un elemento de incertidumbre en los resultados.
- **Retos en asegurar diversidad e independencia:** Si bien se han propuesto métodos para mejorar la diversidad e independencia de los resúmenes, su implementación y eficacia efectiva presentan desafíos.

Para enfrentar estas limitaciones, se han adoptado distintas estrategias. Las estrategias relativas a la generación del dataset fueron implementadas por los autores Stiennon et al. 2020, mientras que las relacionadas con los experimentos fueron llevadas a cabo por los autores de este trabajo:

1. **Aleatoriedad en la generación:** Introducir variabilidad durante el proceso de generación puede producir resúmenes más diversos y limitar similitudes no deseadas.
2. **Diversificación de modelos:** Al usar diferentes configuraciones o arquitecturas, podemos garantizar que los resúmenes no estén sesgados por una única estructura.
3. **Incorporación de fuentes externas:** Añadir resúmenes generados por humanos o por otros modelos puede enriquecer el conjunto de evaluación y proporcionar referencias adicionales.
4. **Mejora en la recopilación de datos:** Un proceso cuidadoso de recopilación y ajuste puede reducir sesgos y dependencias en los resúmenes generados.
5. **Segmentación del dataset:** Para evitar inferencias sesgadas, es crucial segmentar el dataset según la calidad de los resúmenes que lo integran, especialmente cuando se comparan modelos generativos de diferentes complejidades. Así, las conclusiones serán más precisas y robustas si se mantienen consistentes a través de distintas muestras.
6. **Reconocimiento de la subjetividad:** Es esencial equilibrar métricas automáticas con evaluaciones humanas para obtener una visión holística de la calidad de los resúmenes.

Si bien estas estrategias abordan varios de los desafíos presentados, la investigación continua es esencial para mejorar aún más la calidad y fiabilidad de los resúmenes automáticos y su evaluación.

5. Trabajo futuro y Conclusiones

A medida que avanzamos en la frontera de la generación automática de resúmenes, es imperativo reflexionar sobre lo logrado y contemplar los próximos pasos. En este sentido, el presente trabajo ha logrado no solo identificar aspectos clave del estado actual de la investigación, sino también señalar áreas de oportunidad para estudios futuros. En la siguiente sección, retomaremos los puntos centrales de la discusión y exploraremos recomendaciones para futuras investigaciones, basándonos en nuestros hallazgos y las limitaciones observadas (5.1). Además, presentaremos una síntesis reflexiva sobre el impacto y la trascendencia de este estudio en el ámbito del resumen automático y su evaluación (5.2).

5.1. Recomendaciones para Futuras Investigaciones

Las conclusiones de esta investigación abren diversas posibilidades para investigaciones futuras en el campo de la generación automática de resúmenes. A continuación, presentamos algunas recomendaciones esenciales para aquellos investigadores interesados en profundizar en esta área:

- **Inclusión de Métricas Contemporáneas:** Sería beneficioso incluir un mayor número de métricas contemporáneas en el pipeline de la investigación. Esto con el propósito de identificar métricas que estén más alineadas con el juicio humano, y de esta manera, mejorar la precisión del modelo predictivo definido en el método `ExplainSumm`.
- **Caracterización de Resúmenes Escritos por Expertos:** Dada la creciente efectividad de los modelos de generación automática de resúmenes, sería valioso identificar qué distingue a los resúmenes escritos por expertos humanos de aquellos generados automáticamente. Establecer estas diferencias podría ofrecer ideas para optimizar aún más las herramientas automáticas y acercarlas al nivel experto humano.

5.2. Conclusiones Generales

Los avances en la generación automática de resúmenes han llegado a un punto crucial en la investigación contemporánea. A través de múltiples experimentos, este trabajo destaca cómo los sistemas automáticos no solo pueden igualar, sino en ocasiones superar a los resúmenes humanos. Este avance en la calidad pone de manifiesto la imperativa necesidad de reconsiderar y adaptar continuamente nuestras estrategias y métricas de evaluación.

El contraste con investigaciones previas muestra una crítica creciente hacia métricas tradicionales como ROUGE y BLEU. Aunque estas métricas han sido fundamentales en la historia de la generación de resúmenes, su relevancia y precisión están siendo desafiadas en el contexto actual. La

incapacidad de estas métricas para adaptarse al contexto actual, sugiere que se debe avanzar hacia métodos de evaluación más sofisticados y contextuales.

Emergen, por lo tanto, métricas basadas en la similitud del coseno y en embeddings pre-entrenados, las cuales han demostrado una fuerte correlación con el juicio humano. La relevancia de estas métricas, junto con las basadas en la similitud de Jaccard, destaca la importancia de utilizar métodos de evaluación que realmente reflejen la percepción humana en el campo de la generación de resúmenes.

Los resultados también ponen de manifiesto la relevancia del feedback humano en la evaluación y mejora de los sistemas automáticos. El feedback humano no solo sirve como medio de optimización de estos sistemas, sino que también contribuye a la evaluación de los mismos. Integrar la perspectiva humana en la meta-evaluación de métricas de generación automática es crucial para mejorar el alineamiento con el juicio humano.

Por último, `ExplainSumm`, como método innovador propuesto en este trabajo, ha demostrado ser una herramienta valiosa para identificar las métricas y características relevantes en el contexto de generación automática de resúmenes. Su eficiencia y capacidad para detectar patrones complejos lo convierten en una solución prometedora para futuras investigaciones en el campo.

En conjunto, este trabajo enfatiza la necesidad de una perspectiva más holística y centrada en el ser humano para la generación y evaluación automática de resúmenes. Los sistemas automáticos, a medida que avanzan y evolucionan, deben ser evaluados y ajustados en estrecha relación con la percepción y juicio humano. La búsqueda de métricas más alineadas con esta perspectiva será esencial para el futuro de la investigación en generación automática de resúmenes.

REFERENCIAS

Böhm, Florian et al. (nov. de 2019). "Better Rewards Yield Better Summaries: Learning to Summarise Without References". En: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, págs. 3110-3120. DOI: 10.18653/v1/D19-1307. URL: https://aclanthology.org/D19-1307.

Chaganty, Arun, Stephen Mussmann y Percy Liang (jul. de 2018). "The price of debiasing automatic metrics in natural language evalaution". En: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, págs. 643-653. DOI: 10.18653/v1/P18-1060. URL: https://aclanthology.org/P18-1060.

Chopra, Sumit, Michael Auli y Alexander M. Rush (jun. de 2016). "Abstractive Sentence Summarization with Attentive Recurrent Neural Networks". En: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, págs. 93-98. DOI: 10.18653/v1/N16-1012. URL: https://aclanthology.org/N16-1012.

Dong, Yue et al. (oct. de 2018). "BanditSum: Extractive Summarization as a Contextual Bandit". En: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, págs. 3739-3748. DOI: 10.18653/v1/D18-1409. URL: https://aclanthology.org/D18-1409.

Dorr, Bonnie, David Zajic y Richard Schwartz (2003). "Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation". En: *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, págs. 1-8. URL: https://aclanthology.org/W03-0501.

Fabbri, Alexander R. et al. (2020). *SummEval: Re-evaluating Summarization Evaluation*. DOI: 10.48550/ARXIV.2007.12626. URL: https://arxiv.org/abs/2007.12626.

Flesch, R. (1949). *The Art of Readable Writing*. The Art of Readable Writing v. 10. Harper. URL: https://books.google.es/books?id=5RgLAAAAMAAJ.

Hermann, Karl Moritz et al. (2015). "Teaching Machines to Read and Comprehend". En: *Advances in Neural Information Processing Systems*. Ed. por C. Cortes et al. Vol. 28. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/afdec7005cc9f14302cd0474fd0f3c96-Paper.pdf.

Holtzman, Ari et al. (2019). *The Curious Case of Neural Text Degeneration*. DOI: 10.48550/ARXIV.1904.09751. URL: https://arxiv.org/abs/1904.09751.

Jaccard, Paul (ene. de 1901). "Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines." En: *Bulletin de la Societe Vaudoise des Sciences Naturelles* 37, págs. 241-72. DOI: 10.5169/seals-266440.

Kryscinski, Wojciech et al. (nov. de 2019). "Neural Text Summarization: A Critical Evaluation". En: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, págs. 540-551. DOI: 10.18653/v1/D19-1051. URL: https://aclanthology.org/D19-1051.

Lavie, Alon y Abhaya Agarwal (2007). "Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments". En: *Proceedings of the Second Workshop on Statistical Machine Translation*. StatMT '07. Prague, Czech Republic: Association for Computational Linguistics, págs. 228-231.

Leike, Jan et al. (2018). *Scalable agent alignment via reward modeling: a research direction*. DOI: 10.48550/ARXIV.1811.07871. URL: https://arxiv.org/abs/1811.07871.

Li, Margaret, Jason Weston y Stephen Roller (2019). *ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons*. DOI: 10.48550/ARXIV.1909.03087. URL: https://arxiv.org/abs/1909.03087.

Lin, Chin-Yew (jul. de 2004). "ROUGE: A Package for Automatic Evaluation of Summaries". En: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, págs. 74-81. URL: https://aclanthology.org/W04-1013.

Maynez, Joshua et al. (jul. de 2020). "On Faithfulness and Factuality in Abstractive Summarization". En: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, págs. 1906-1919. DOI: 10.18653/v1/2020.acl-main.173. URL: https://aclanthology.org/2020.acl-main.173.

Nallapati, Ramesh, Feifei Zhai y Bowen Zhou (2017). "SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents". En: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, págs. 3075-3081.

Nallapati, Ramesh et al. (ago. de 2016). "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond". En: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, págs. 280-290. DOI: 10.18653/v1/K16-1028. URL: https://aclanthology.org/K16-1028.

Narayan, Shashi, Shay B. Cohen y Mirella Lapata (jun. de 2018). "Ranking Sentences for Extractive Summarization with Reinforcement Learning". En: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, págs. 1747-1759. DOI: 10.18653/v1/N18-1158. URL: https://aclanthology.org/N18-1158.

Novikova, Jekaterina et al. (sep. de 2017). "Why We Need New Evaluation Metrics for NLG". En: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, págs. 2241-2252. DOI: 10.18653/v1/D17-1238. URL: https://aclanthology.org/D17-1238.

Papineni, Kishore et al. (jul. de 2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". En: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, págs. 311-318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040.

Paulus, Romain, Caiming Xiong y Richard Socher (2017). *A Deep Reinforced Model for Abstractive Summarization*. DOI: 10.48550/ARXIV.1705.04304. URL: https://arxiv.org/abs/1705.04304.

Radford, Alec y Karthik Narasimhan (2018). "Improving Language Understanding by Generative Pre-Training". En: URL: https://api.semanticscholar.org/CorpusID:49313245.

Reimers, Nils e Iryna Gurevych (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. DOI: 10.48550/ARXIV.1908.10084. URL: https://arxiv.org/abs/1908.10084.

Ross, Stephane, Geoffrey J. Gordon y J. Andrew Bagnell (2010). *A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning*. DOI: 10.48550/ARXIV.1011.0686. URL: https://arxiv.org/abs/1011.0686.

Rush, Alexander M., Sumit Chopra y Jason Weston (sep. de 2015). "A Neural Attention Model for Abstractive Sentence Summarization". En: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, págs. 379-389. DOI: 10.18653/v1/D15-1044. URL: https://aclanthology.org/D15-1044.

Schluter, Natalie (abr. de 2017). "The limits of automatic summarisation according to ROUGE". En: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, págs. 41-45. URL: https://aclanthology.org/E17-2007.

Schmidt, Florian (2019). *Generalization in Generation: A closer look at Exposure Bias*. DOI: 10.48550/ARXIV.1910.00292. URL: https://arxiv.org/abs/1910.00292.

Schulman, John et al. (2017). *Proximal Policy Optimization Algorithms*. DOI: 10.48550/ARXIV.1707.06347. URL: https://arxiv.org/abs/1707.06347.

See, Abigail, Peter J. Liu y Christopher D. Manning (jul. de 2017). "Get To The Point: Summarization with Pointer-Generator Networks". En: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, págs. 1073-1083. DOI: 10.18653/v1/P17-1099. URL: https://aclanthology.org/P17-1099.

Stiennon, Nisan et al. (2020). *Learning to summarize from human feedback*. DOI: 10.48550/ARXIV.2009.01325. URL: https://arxiv.org/abs/2009.01325.

Touvron, Hugo et al. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. DOI: 10.48550/ARXIV.2307.09288. URL: https://arxiv.org/abs/2307.09288.

Völske, Michael et al. (2017). "TL;DR: Mining Reddit to Learn Automatic Summarization". En: *Proceedings of the Workshop on New Frontiers in Summarization*. Association for Computational Linguistics. DOI: 10.18653/v1/w17-4508. URL: https://doi.org/10.18653/v1/w17-4508.

Wu, Yuxiang y Baotian Hu (abr. de 2018). "Learning to Extract Coherent Summary via Deep Reinforcement Learning". En: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1. DOI: 10.1609/aaai.v32i1.11987. URL: https://doi.org/10.1609/aaai.v32i1.11987.

Zhang, Hugh et al. (2020). *Trading Off Diversity and Quality in Natural Language Generation*. DOI: 10.48550/ARXIV.2004.10450. URL: https://arxiv.org/abs/2004.10450.

Zhang, Tianyi et al. (2019). *BERTScore: Evaluating Text Generation with BERT*. DOI: 10.48550/ARXIV.1904.09675. URL: https://arxiv.org/abs/1904.09675.

Ziegler, Daniel M. et al. (2019). *Fine-Tuning Language Models from Human Preferences*. DOI: 10.48550/ARXIV.1909.08593. URL: https://arxiv.org/abs/1909.08593.

APPENDICES

A. DEFINICIÓN DE MÉTRICAS Y RASGOS

La evaluación cuantitativa de los resúmenes es un paso esencial para determinar la efectividad de los métodos de resumen automático. En la Sección 3.2.2, se introducen métricas y rasgos clave en esta evaluación.

Las métricas tradicionales como ROUGE y BLEU, se enfocan en la coincidencia exacta de palabras o n-gramas entre el texto fuente y el resumen. Sin embargo, con la evolución de las técnicas de procesamiento de lenguaje natural, la necesidad de considerar la similitud semántica ha llevado al desarrollo y adopción de métricas basadas en embeddings como la Similitud del Coseno. Además, la Similitud de Jaccard brinda una perspectiva sobre el solapamiento entre conjuntos de términos, permitiendo una comprensión más amplia de la similitud entre textos.

Por otro lado, no basta con que un resumen sea similar al original; también debe ser legible y coherente. Es aquí donde entran en juego los rasgos de legibilidad y complejidad lingüística, que ofrecen información sobre la estructura y construcción del resumen desde un punto de vista lingüístico.

A continuación, ofrecemos una descripción detallada de estas métricas y rasgos, su propósito y las fórmulas que las rigen, permitiendo una comprensión profunda de cómo cada una contribuye a la evaluación integral de los resúmenes generados.

A.1. MÉTRICAS ROUGE

El método ROUGE (Recall-Oriented Understudy for Gisting Evaluation) fue propuesto por el autorLin 2004, y su principal objetivo es evaluar la calidad de los resúmenes comparándolos con resúmenes de referencia utilizando n-gramas.

ROUGE-1. Esta métrica compara la coincidencia de unigramas (palabras) entre los resúmenes generados y los resúmenes de referencia. ROUGE-1 se calcula usando la fórmula:

$$\text{ROUGE-1} = \frac{\sum_{i \in \text{resúmenes generados}} \sum_{j \in \text{resúmenes referencia}} \delta(i = j)}{\sum_{j \in \text{resúmenes referencia}} 1}$$

Donde:

- $\delta(x)$: es una función indicadora que es igual a 1 si x es verdadero y 0 en caso contrario.

ROUGE-2. Centrándose en bigramas, ROUGE-2 extiende la métrica anterior para considerar pares de palabras consecutivas. La fórmula, aunque similar a ROUGE-1, se enfoca en bigramas en vez de palabras individuales:

$$\text{ROUGE-2} = \frac{\sum_{i \in \text{bigramas generados}} \sum_{j \in \text{bigramas referencia}} \delta(i = j)}{\sum_{j \in \text{bigramas referencia}} 1}$$

ROUGE-L. Diferente a las anteriores, ROUGE-L se basa en la longitud de la subsecuencia común más larga (LCS) entre los resúmenes generados y de referencia. Su cálculo se expresa mediante la fórmula:

$$\text{ROUGE-L} = \frac{\text{LCS}(\text{resumen generado}, \text{resumen referencia})}{\text{longitud del resumen de referencia}}$$

A.2. MÉTRICA BLEU

La métrica BLEU (Bilingual Evaluation Understudy) fue diseñada principalmente para evaluar la calidad de las traducciones automáticas en comparación con traducciones humanas de referencia Papineni et al. 2002. Sin embargo, también ha encontrado utilidad en tareas de generación de texto como resúmenes. La fórmula general es:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log(p_n)\right)$$

Donde:

- BP: es la penalización por brevedad, que desfavorece las respuestas cortas. Se calcula como:
$$BP = \text{mín}\left(1, e^{(1-\frac{\text{longitud del resumen de referencia}}{\text{longitud del resumen generado}})}\right)$$
- p_n: es la precisión de los n-gramas.
- w_n: son los pesos asignados a cada p_n, que generalmente se establecen para que su suma sea 1.
- N: es el máximo tamaño de los n-gramas a considerar, que generalmente es 4.

A.3. SIMILITUD DE JACCARD

La Similitud de Jaccard, también conocida como el índice de Jaccard o coeficiente de similitud de Jaccard, evalúa el solapamiento de palabras o n-gramas entre dos conjuntos de texto. Este coeficiente fue introducido por el autor Jaccard 1901 como la relación entre el tamaño de la intersección de dos conjuntos y el tamaño de su unión.

La fórmula es:
$$\text{Similitud de Jaccard} = \frac{|\text{conjunto A} \cap \text{conjunto B}|}{|\text{conjunto A} \cup \text{conjunto B}|}$$

Donde:

- conjunto A: representa las palabras o n-gramas en el texto original.
- conjunto B: representa las palabras o n-gramas en el resumen.

Para n-gramas de tamaño 1, se consideran palabras individuales, mientras que para tamaño 2 se consideran bigramas.

A.4. SIMILITUD DEL COSENO

La Similitud del Coseno mide el coseno del ángulo entre dos vectores no nulos. En este contexto, los vectores pueden ser los embeddings del texto original o del resumen de referencia en comparación con el embedding del resumen generado. Esta métrica devuelve un valor entre -1 y 1, donde 1 indica una similitud total, 0 indica que los vectores son ortogonales (no relacionados), y -1 indica que los vectores son diametralmente opuestos. Para los embeddings, un valor más cercano a 1 indica una mayor similitud en términos de contenido semántico entre el texto original (o resumen de referencia) y el resumen.

La fórmula es:
$$\text{Similitud del Coseno} = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}||_2 \times ||\mathbf{B}||_2}$$

Donde:

- $\mathbf{A}$: es el vector embedding del texto original o del resumen de referencia.
- $\mathbf{B}$: es el vector embedding del resumen generado.

Respecto a la generación de embeddings, se utiliza la librería SentenceTransformers en Python, un framework especializado en la generación de embeddings de oraciones, texto e imágenes, y que se describe detalladamente en el artículo ”Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks” de Reimers et al. 2019. En concreto, para proyectar estos embeddings, se ha seleccionado el modelo all-MiniLM-L6-v2. Este modelo tiene la capacidad de mapear oraciones y párrafos a un espacio vectorial denso de 384 dimensiones, lo que lo hace particularmente útil para tareas como la búsqueda o agrupación semántica.

El enfoque aquí es trascender más allá de la coincidencia exacta de palabras y capturar la esencia semántica de los textos. Una vez generados, estos embeddings se normalizan para asegurar que todos los vectores tengan una longitud unitaria. Esta normalización es crucial antes de calcular la similitud del coseno entre los embeddings, ya que permite que la métrica refleje de manera precisa el grado de similitud semántica entre dos textos, independientemente de su longitud.

A.5. Ratio de Compresión

El Ratio de Compresión es una métrica que indica cuánto ha sido condensado el texto original para producir el resumen. Un ratio más bajo indica un resumen más corto en comparación con el texto original.

La fórmula es:

$$\text{Ratio de Compresión} = \frac{\text{longitud del resumen}}{\text{longitud del texto original}}$$

A.6. Rasgos de Legibilidad y Complejidad Lingüística

Estos rasgos se centran en analizar cómo se presenta el resumen en términos de accesibilidad y complejidad lingüística. Los elementos evaluados proporcionan una visión detallada de la estructura y construcción del resumen.

Recuentos Específicos. Los siguientes recuentos aportan información detallada sobre la composición del resumen:

- **Recuento de Sílabas:** Número de sílabas presentes en el resumen.
- **Recuento de Léxicos:** Cantidad de palabras únicas en el resumen.
- **Recuento de Oraciones:** Número total de oraciones en el resumen.
- **Recuento de Caracteres:** Número total de caracteres presentes.
- **Recuento de Letras:** Cantidad total de letras en el resumen.
- **Recuento de Polílabos:** Palabras que tienen tres sílabas o más.
- **Recuento de Monosílabos:** Palabras que constan de una sola sílaba.

Flesch Reading Ease Score (RE). Utilizando los rasgos lingüísticos mencionados, se puede calcular el Flesch Reading Ease Score (RE), una métrica cuantitativa que evalúa la facilidad de lectura de un texto. El índice RE, desarrollado por el autor Flesch 1949 en su libro titulado "The Art of Readable Writing", frece una perspectiva sobre cuán accesible y legible es el resumen para el lector final. La fórmula general para calcular RE es:

$$\text{RE} = 206{,}835 - (1{,}015 \times \frac{\text{Total de palabras}}{\text{Total de oraciones}}) - (84{,}6 \times \frac{\text{Total de sílabas}}{\text{Total de palabras}})$$

Un valor más alto de RE indica un texto más fácil de leer, mientras que un valor más bajo sugiere un texto más complejo.

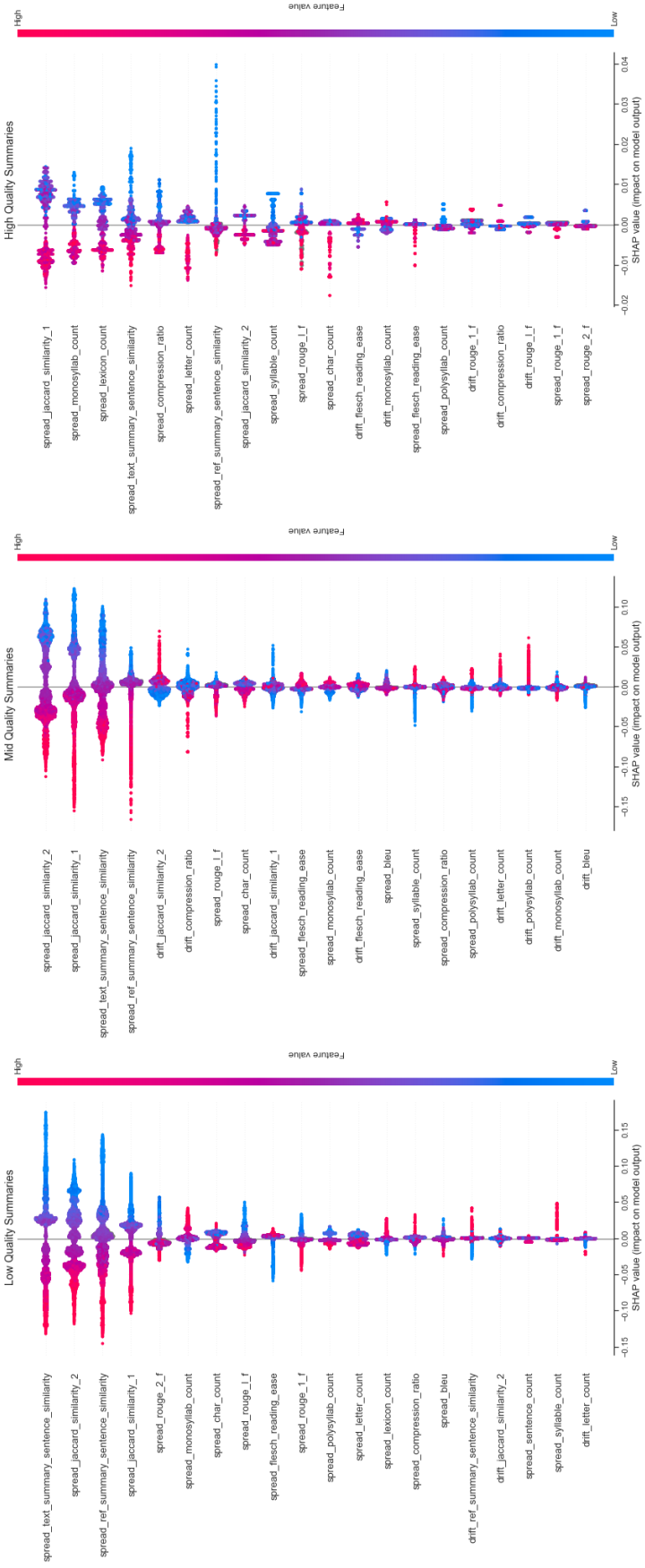

Figura 8: Valores SHAP en cada uno de los 3 subconjuntos.

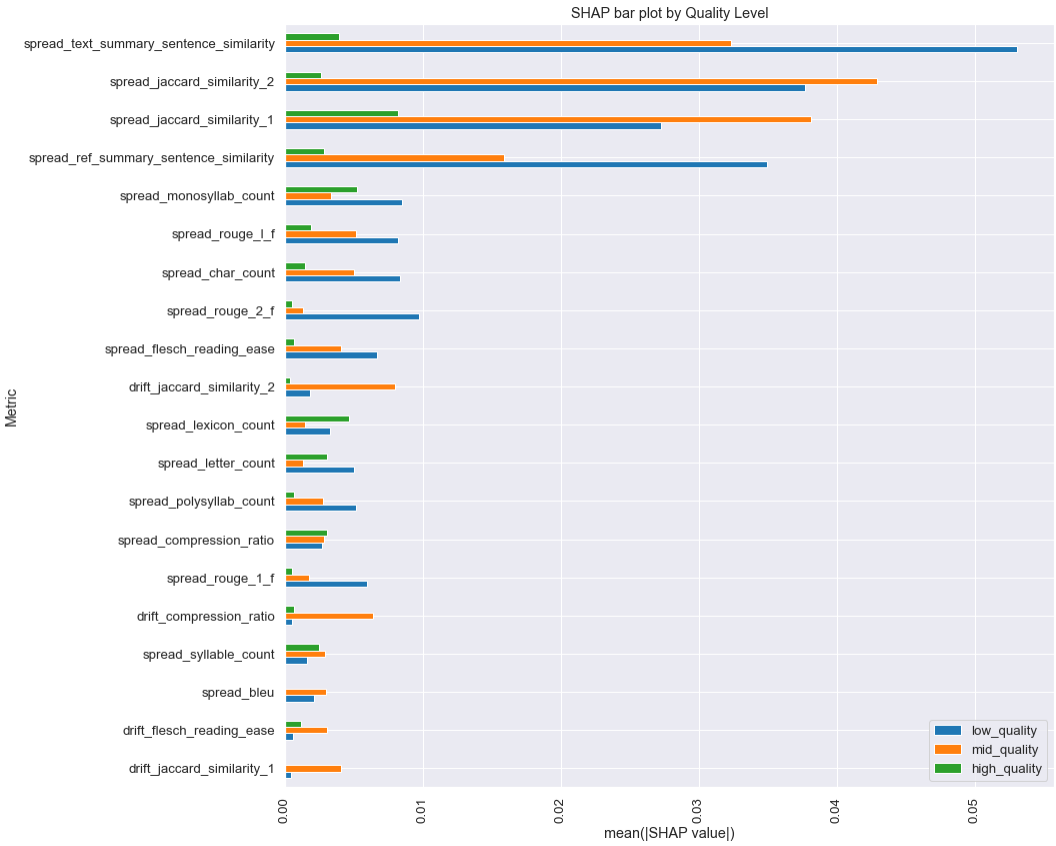

Figura 9: Media de los valores absoluto de los valores SHAP para cada una las características del modelo de cada uno de los 3 subconjuntos. Las métricas están ordenadas de mayor a menor importancia.

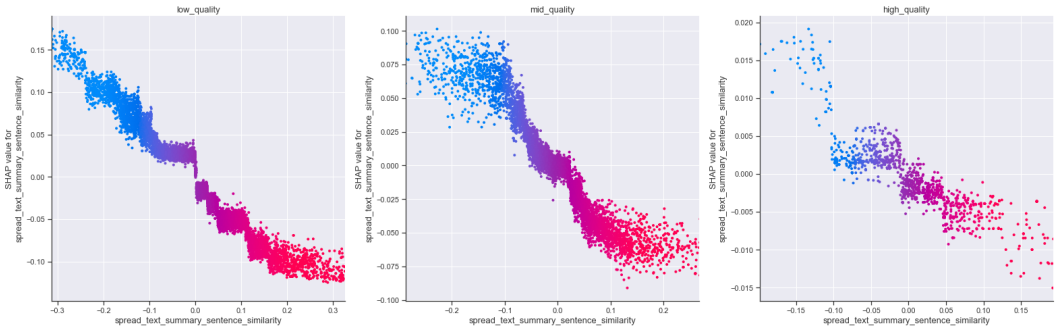

Figura 10: Gráfico de dependencia SHAP para la similitud del coseno entre los embeddings del texto y el resumen utilizando la librería SentenceTransformer.

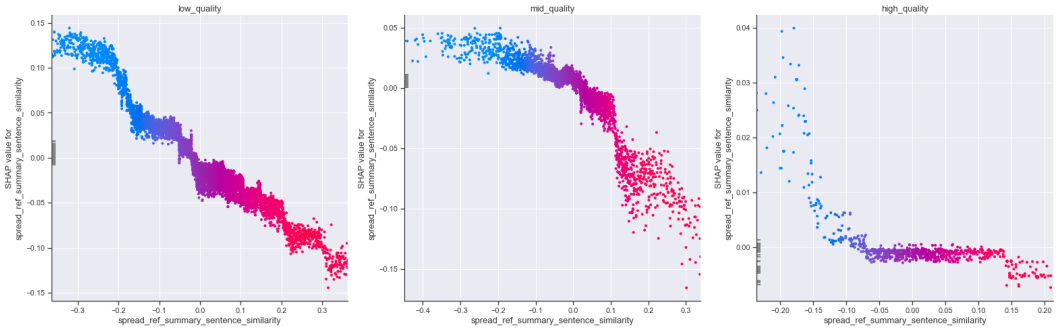

Figura 11: Gráfico de dependencia SHAP para la similitud del coseno entre los embeddings del resumen de referencia y el resumen dado utilizando la librería SentenceTransformer.

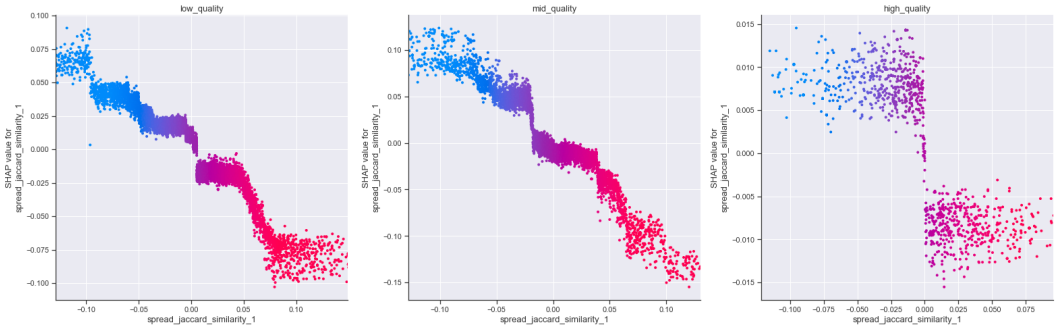

Figura 12: Gráfico de dependencia SHAP para el Coeficiente de Jaccard basado en 1-gramas en el modelo definido.

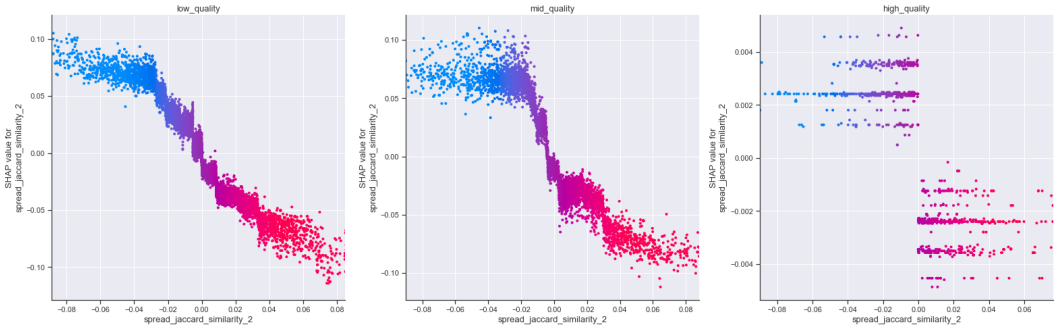

Figura 13: Gráfico de dependencia SHAP para el Coeficiente de Jaccard basado en 2-gramas en el modelo definido.

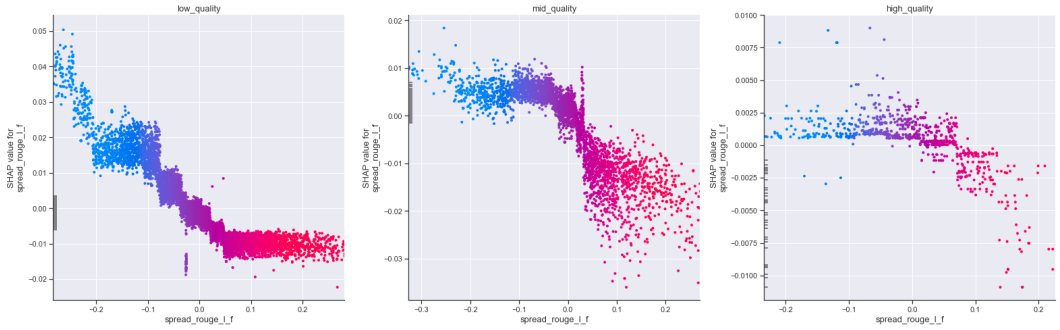

Figura 14: Gráfico de dependencia SHAP para la métrica ROUGE-L F1 en el modelo definido.

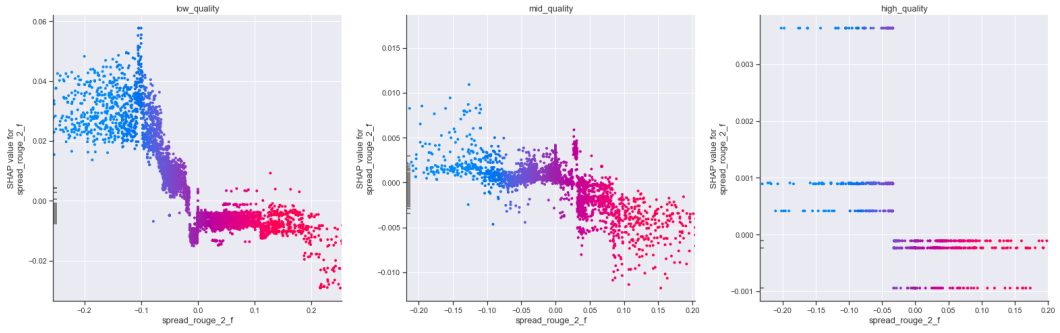

Figura 15: Gráfico de dependencia SHAP para la métrica ROUGE-2 F1 en el modelo definido.

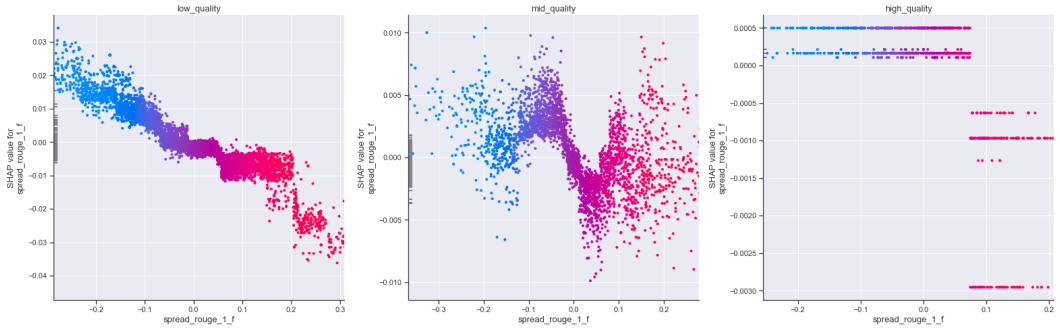

Figura 16: Gráfico de dependencia SHAP para la métrica ROUGE-1 F1 en el modelo definido.

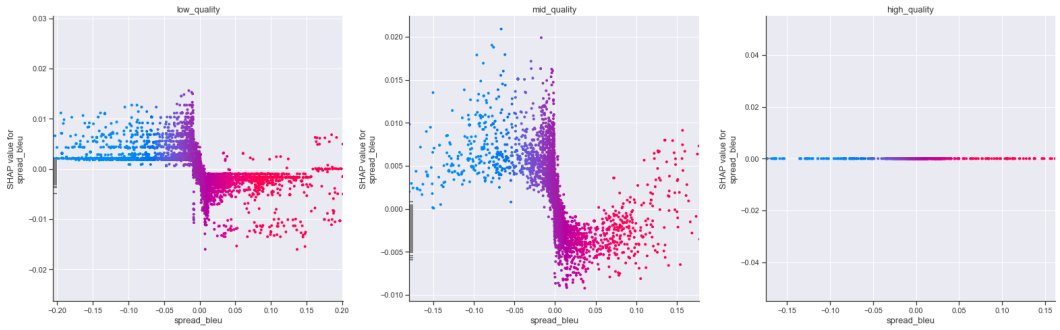

Figura 17: Gráfico de dependencia SHAP para la métrica BLEU en el modelo definido.

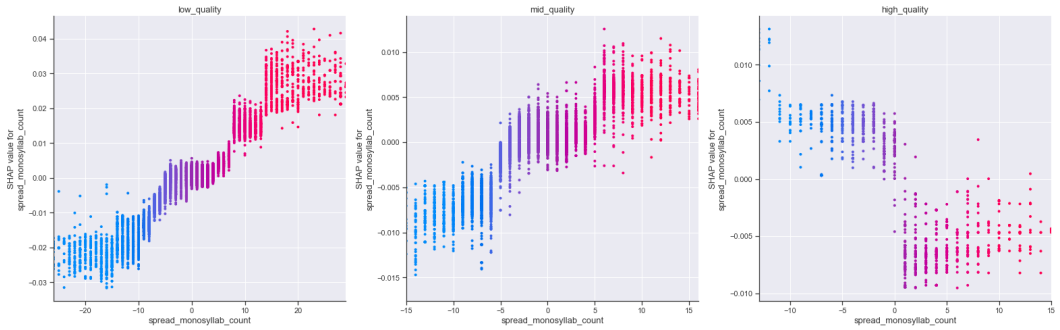

Figura 18: Gráfico de dependencia SHAP para el rasgo ”Número de palabras monosílabas (Monosyllable Count).[en] el modelo definido.

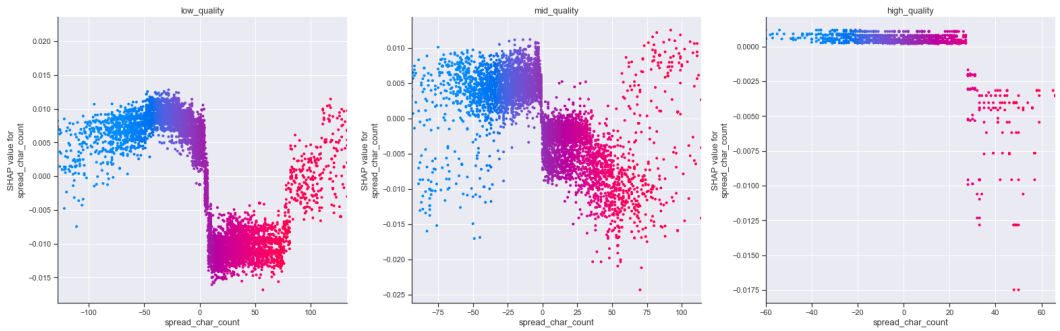

Figura 19: Gráfico de dependencia SHAP para el rasgo ”Número de caracteres (Character Count).[en] el modelo definido.

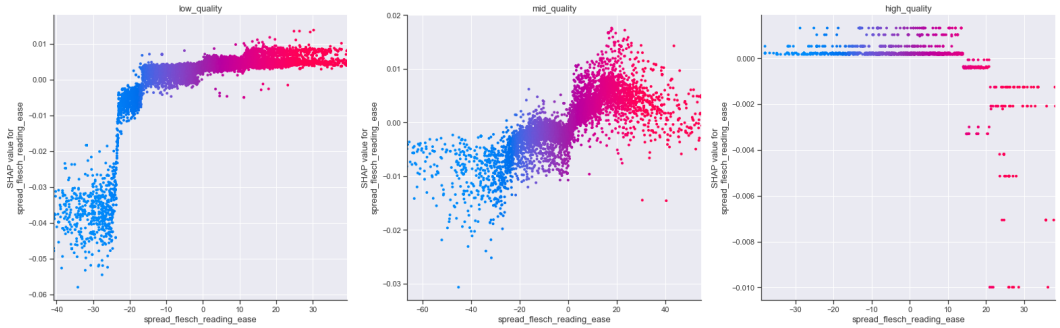

Figura 20: Gráfico de dependencia SHAP para la métrica ”Puntuación de facilidad de lectura de Flesch (Flesch Reading Ease score).[en] el modelo definido.

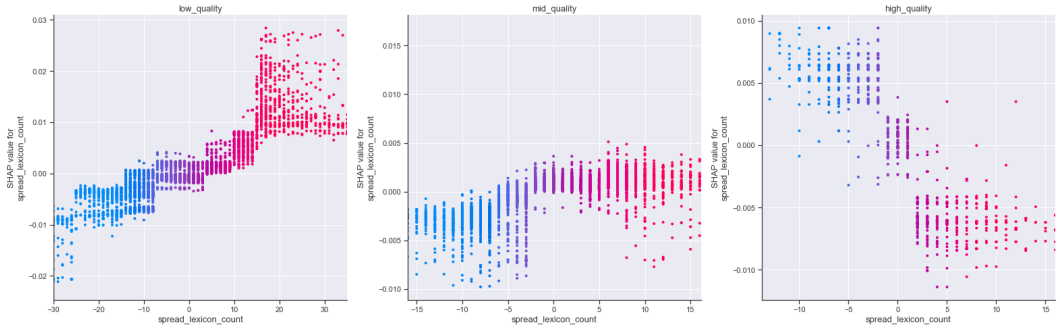

Figura 21: Gráfico de dependencia SHAP para el rasgo Çonteo de palabras (Lexicon Count).[en] el modelo definido.