GENERACIÓN AUTOMÁTICA DE RESÚMENES EN PORTUGUÉS UTILIZANDO ALGORITMOS GENÉTICOS

Griselda Areli Matías Mendoza, Yulia Ledeneva y René Arnulfo García-Hernández
Universidad Autónoma del Estado de México, México

RESUMEN

El lenguaje portugués, hablado por más de 260 millones de personas en países como Brasil, Portugal y varias naciones africanas, juega un papel importante en la sociedad del conocimiento global, por esta razón es importante poder acceder a la información de los diferentes documentos digitales de forma rápida y sin perder la idea principal del documento, para ello se sugiere el uso de la generación automática de resúmenes. Sin embargo, el número de investigaciones sobre la generación automática de resúmenes en este lenguaje es limitado. Este trabajo resalta la importancia de desarrollar métodos efectivos que optimicen la selección de oraciones relevantes utilizando técnicas como los algoritmos genéticos y modelos de texto avanzados como la Bolsa de Palabras (BoW) y los n-gramas. Para este trabajo, se utilizó el corpus TeMário, que consiste en 100 textos periodísticos con resúmenes generados por expertos humanos, y se implementó la métrica ROUGE para evaluar la calidad de los resúmenes automáticos generados. El enfoque principal se basó en la técnica de resúmenes extractivos.

PALABRAS CLAVE

Resúmenes Extractivos, Algoritmos Genéticos, Lenguaje Portugués

1. INTRODUCCIÓN

En la actualidad, el acceso a internet y el uso de las tecnologías de la información ha transformado de manera profunda la dinámica de las sociedades globales, permitiendo una amplia difusión de información digital. Esta tendencia no solo ha incrementado el volumen de datos disponibles, sino que también ha ampliado el alcance de diversas lenguas, entre ellas el portugués.

El portugués, es hablado por más de 260 millones de personas en países como Brasil, Portugal y varias naciones africanas, el portugués se ha consolidado como un lenguaje clave en la sociedad del conocimiento. Su relevancia en este contexto no solo radica en su extensión geográfica y demográfica, sino en su capacidad para generar valor económico y facilitar la comunicación en mercados emergentes. La expansión del portugués en el ámbito de la tecnología y la información, así como su creciente presencia en internet, refuerza su papel como un medio de conexión y un activo esencial en la globalización del conocimiento (Albuquerque, 2010).

Una forma de divulgar la información es por medio de noticias, las noticias tienen como objetivo informar a las audiencias sobre eventos relevantes que ocurren en la sociedad (Califano, 2015). Cada día se generan cientos de noticias en diferentes lenguajes, lo que provoca una sobrecarga de información. Este fenómeno afecta la capacidad de los usuarios para procesar todo el contenido disponible y obtener de forma rápida y eficiente la información más relevante. En este contexto, el Procesamiento del Lenguaje Natural (PLN) aplicado a la Generación Automática de Resúmenes ha emergido como una herramienta clave para contrarrestar los efectos de la sobrecarga informativa.

Según Supriyono (2024), la generación automática de resúmenes es una técnica fundamental dentro del PLN que permite convertir grandes cantidades de datos textuales en representaciones más concisas y comprensibles. Esta técnica implica extraer la información más relevante de un documento o conjunto de documentos para proporcionar una versión más breve, pero representativa, del contenido original (Supriyono et al., 2024). Existen tres enfoques principales en la generación automática de resúmenes según (El-Kassas et al., 2021) y (Neri-Mendoza et al., 2024):

1. Resúmenes Extractivos: Este enfoque selecciona las oraciones o fragmentos clave del texto original y los combina para formar el resumen. No se requiere una reinterpretación del contenido, lo que lo convierte en un proceso más directo y basado en la identificación de las partes más relevantes del texto.
2. Resúmenes Abstractivos: A diferencia del extractivo, este enfoque genera nuevas frases que no necesariamente aparecen en el texto original. Los resúmenes abstractivos implican una mayor comprensión del contenido, ya que el modelo debe parafrasear o reescribir el texto en sus propias palabras, lo que requiere un nivel más profundo de procesamiento del lenguaje.
3. Resúmenes Híbridos: Combinan ambos enfoques, utilizando elementos de la selección directa de oraciones y la reescritura del contenido para generar un resumen más coherente y preciso.

En este artículo nos enfocamos en la generación de resúmenes extractivos, utilizando técnicas de optimización, las cuales nos permiten extraer las oraciones de forma más sencilla, con los mejores resultados del estado de arte y la idea principal del texto.

Los resúmenes extractivos son de gran utilidad no solo para conocer la idea principal de un texto, sino también como complemento de otras tareas como la clasificación de documentos (Isonuma et al., 2017), detección de noticias falsas (Garcia et al., 2024), entre otras.

Figura 1. Ejemplo de resumen extractivo

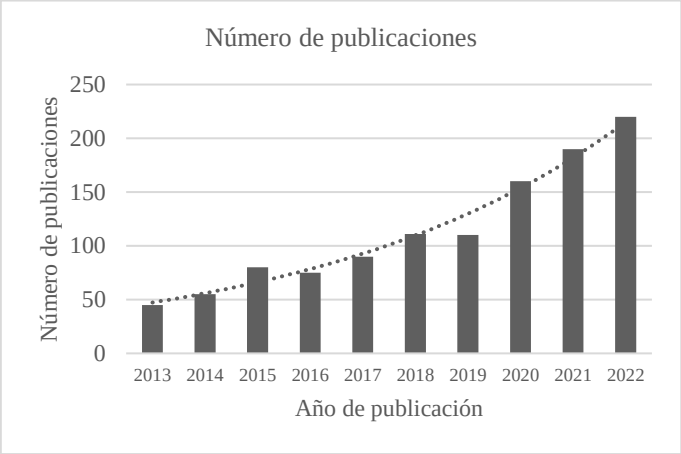
Para la investigación de la generación automática de resúmenes, es necesario contar con los recursos necesarios como son el corpus estándar y la métrica de evaluación, esto permite poder probar y medir los diferentes métodos del estado del arte y las diferentes herramientas que generan resúmenes. Para el lenguaje portugués se tiene un corpus estándar de noticias, llamado TeMário (Pardo and Rino, 2003). Utilizando el corpus se pueden realizar pruebas, las cuales deben ser evaluadas con una herramienta, en nuestro caso usaremos ROUGE (Lin, 2004) la cual permite comparar los resúmenes generados a partir de los métodos con los resúmenes generados por un humano.

Existen numerosas investigaciones que aplican algoritmos de optimización, en su mayoría centradas en el lenguaje inglés. Este trabajo emplea un algoritmo de optimización, el cual permite abordar problemas con una gran cantidad de soluciones posibles, como la selección óptima de oraciones en un conjunto de texto para generar resúmenes precisos.

Aunque la mayoría de las investigaciones de generación automática de resúmenes se enfocan en inglés, lenguajes como el portugués, han recibido menor atención. Esto resalta la necesidad de desarrollar métodos para la generación automática de resúmenes en portugués, que puedan atender las particularidades lingüísticas y ampliar el alcance de la tecnología de PLN en contextos de habla portuguesa.

2. ESTADO DEL ARTE

Como ya se mencionó la mayor parte de la investigación está en lenguaje inglés. Según Jin (2024) el avance de las investigaciones sobre la tarea de generación automática de resúmenes en los últimos años ha ido en aumento, por lo que haciendo un análisis del crecimiento se estima que las investigaciones crecen en un 20% anual. Sin embargo, en su gran mayoría son en el lenguaje inglés.



Gráfica 1. Número de publicaciones en Generación Automática de Resúmenes de 2013 a 2022, *(Jin et al., 2024)*

Para portugués, el número de investigaciones es menor, ya que por año se realizan entre una y dos investigaciones, esto hace que en la tarea de generación automática de resúmenes para este lenguaje tenga una amplia oportunidad de investigación.

A continuación, se describen las investigaciones más relevantes sobre la generación automática de resúmenes en portugués. La Tabla 1 muestra los principales trabajos realizados en esta área, donde se observa que existen muy pocos corpus estándar para el lenguaje, siendo el corpus TeMário el más utilizado. El tipo de resumen que más se realiza es el de tipo extractivo.

Tabla 1. Estado del arte de la generación automática de resúmenes en portugués

Artículo	Tipo de Resumen	Corpus Utilizado	Lenguaje	Métrica de Evaluación	Técnica de PLN Utilizada
(Pardo and Rino, 2003)	Extractivo	TeMário	Portugués	No especificada	Creación del corpus TeMário03
(Rino et al., 2004)	Comparación	TeMário	Portugués	Precisión, Recall	Comparación de métodos automáticos
(Mihalcea and Tarau, 2005)	Multidocumento	Corpus multilingüe	Portugués, inglés	ROUGE	Grafos y algoritmos basados en grafos
(Orrú et al., 2006)	Extractivo	No aplica	Portugués	No especificada	Redes neuronales artificiales
(Pardo et al., 2006)	Evaluación de resúmenes	TeMário	Portugués	ROUGE, métricas de redes	Redes complejas
(Maziero et al., 2007)	Extractivo	TeMário	Portugués	No especificada	Creación del corpus TeMário06
(Antiqueira, 2007)	Extractivo	TeMário	Portugués	ROUGE	Redes complejas
(Margarido et al 2008)	Simplificación	Lectores funcionalmente analfabetos	Portugués	Comprensión del lector	Simplificación textual
(Nunes et al., 2010)	No aplica	No aplica	Portugués	No aplica	Análisis computacional del lenguaje. PLN en Brasil

(de Oliveira and Guelpeli, 2011)	Extractivo	Varios textos de prueba	Portugués	ROUGE	Metaheurísticas de optimización. Modelo BLMSumm basado en búsqueda local
(Cardoso et al., 2011)	Multidocumento	Corpus CSTNews (50 clusters de noticias en portugués brasileño)	Portugués brasileño	*F-measure*, Precisión, *Recall*	Modelos de estructura retórica y cruzada. Creación de un corpus anotado para la sumarización de noticias
(Amancio et al., 2012)	Extractivo	TeMário	Portugués	ROUGE	Redes complejas
(Oliveira and Guelpeli, 2012)	Multilingüe	Varios textos de prueba	Múltiples lenguajes	ROUGE	Evaluación de múltiples lenguajes con métodos estadísticos
(Cabral et al., 2014)	Extractivo	Varios corpus multilingües	Multilingüe	ROUGE	Algoritmos independientes del lenguaje
(Cavalieri et al., 2015)	Predicción de palabras	Varios corpus	Portugués, inglés, español	Precisión, KSS	Modelos de lenguaje con N-*grams*
(Asevedo Nóbrega and Salgueiro Pardo 2017)	Resumen de actualización	Colección de textos en portugués	Portugués	ROUGE, relevancia	Métodos combinados
(Silva and Pardo, 2022)	Resúmenes contrastivos	Conjunto de opiniones	Portugués/inglés	Representatividad, diversidad	Heurística basada en opiniones
(Paiola et al., 2022)	Resumen abstractivo	Corpus en portugués brasileño	Portugués brasileño	Coherencia, ROUGE	Deep learning
(Lins et al., 2024)	Resumen abstractivo	Ruling.BR	Portugués	ROUGE (ROUGE-1, ROUGE-2, ROUGE-L)	Modelos de lenguaje profundo, BERTimbau, técnicas extractivas
(Mussandi and Wichert, 2024)	No aplica	Lenguas africanas	Multilengujes	No aplica	No aplica. Análisis de lenguas con pocos recursos

3. MÉTODO PROPUESTO

En esta sección, se presenta el método propuesto para la generación automática de resúmenes en portugués, dividido en varias etapas clave para optimizar la calidad del resumen generado. Primero, se describe el corpus utilizado en el experimento, que servirá como base de datos para la extracción de contenido relevante. Luego, se detalla la fase de preprocesamiento, que incluye técnicas para limpiar y estructurar el texto antes de aplicar el modelo. A continuación, se expone el uso de un algoritmo genético como estrategia de optimización, adaptado para seleccionar las oraciones clave del texto de manera eficiente. Posteriormente, se explica la construcción del resumen a partir de las oraciones seleccionadas, seguida de las pruebas experimentales realizadas para verificar la efectividad del método. Finalmente, se discute la evaluación del resumen generado mediante métricas establecidas, lo que permite medir la calidad del resumen y compararlo con resúmenes de referencia.

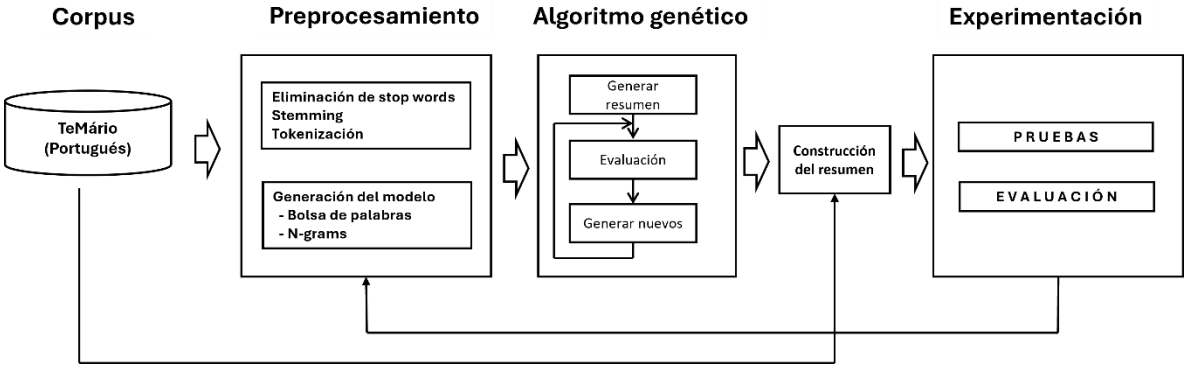


Figura 2. Diagrama del método propuesto

3.1 Corpus

El corpus estándar por utilizar es TeMário. El corpus TeMário (TExtos com suMÁRIOS) es un recurso creado específicamente para la investigación en generación automática de resúmenes en portugués. Está compuesto por 100 textos periodísticos extraídos de dos periódicos brasileños, Folha de São Paulo y Jornal do Brasil, distribuidos en cinco categorías temáticas: especial, mundo, opinión, política e internacional. Contiene un resumen para cada texto el cual ha sido resumido por un experto humano, generado de entre el 25% y el 30% de la longitud del documento original. Considerando la longitud que maneja el corpus para las pruebas de este trabajo se consideró el 30%.

3.2 Preprocesamiento

La etapa de preprocesamiento es fundamental para optimizar el rendimiento de los modelos de generación automática de resúmenes. Esta fase tiene como objetivo preparar el texto, eliminando información irrelevante y normalizando las palabras para mejorar la eficiencia de los algoritmos de resumen. Para la etapa de preprocesamiento, hicimos uso de la eliminación de stopwords, la cual consiste en eliminar las palabras que no aportan valor semántico en el caso del lenguaje portugués, este proceso es crucial para reducir el ruido que pueden generar las palabras comunes y mejorar el rendimiento de los algoritmos de generación automática de resúmenes. Algunos ejemplos son artículos como o, a, os, as, um, uma, y preposiciones como de, do, da, no, na, em, por, com.

También realizamos experimentos utilizando dos tipos de tokenización: por oraciones y por párrafos. primero utilizamos tokenización por oraciones, ya que trabajar con oraciones permite preservar la estructura gramatical de cada oración y es útil para métodos extractivos que seleccionan oraciones completas como resúmenes. Además, trabajamos con tokenización por párrafos, ya que esta permite tener una visión más amplia de las ideas del texto, lo que resulta útil en tareas que requieren analizar la coherencia entre varias oraciones, como en resúmenes abstractivos.

Consideramos que probar ambos enfoques es esencial para determinar cuál ofrece mejor balance entre concisión y coherencia, en función del método de generación automática de resúmenes utilizado.

Además del preprocesamiento, es fundamental definir el modelo de texto con el que se trabajara. El modelo de texto describe cómo se representará la información dentro del texto, influyendo directamente en el análisis y el resumen resultante. Los modelos de texto considerados en este trabajo son: Bolsa de palabras (BoW): Este modelo representa el texto como un conjunto de palabras, sin tener en cuenta su orden. N-gramas: Los n-gramas representan secuencias de palabras consecutivas, donde "n" indica el número de palabras en la secuencia. Probar con diferentes tamaños de n-gramas (bi-gramas, tri-gramas, cuatri-gramas) permite capturar relaciones contextuales entre palabras, lo cual es crucial para conservar la coherencia semántica en el resumen, por lo que en este trabajo se hace uso de bi-gramas, tri-gramas y cuatri-gramas.

3.3 Algoritmo Genético

Los algoritmos genéticos destacan como una solución eficiente en la generación automática de resúmenes, al abordar problemas complejos que requieren seleccionar opciones óptimas entre numerosas combinaciones posibles (combinación de oraciones). En el caso de los resúmenes extractivos, el desafío consiste en identificar las oraciones más relevantes de un documento para construir un resumen coherente y representativo. Este proceso, al implicar múltiples configuraciones potenciales, se plantea como un problema de optimización que los algoritmos genéticos resuelven al explorar de manera eficaz el espacio de soluciones posibles. A continuación, se describe el algoritmo utilizado.

Para la codificación del cromosoma, se emplea una representación binaria que refleja la selección o exclusión de oraciones del texto original. En esta representación, cada gen del cromosoma corresponde a una oración específica: si el valor del gen es 1, significa que la oración asociada ha sido seleccionada para formar parte del resumen; en cambio, si el valor del gen es 0, indica que dicha oración no será incluida en el resumen.

La población inicial está constituida por un conjunto de cromosomas, cada uno representando un posible resumen del texto. Estos cromosomas se generan de manera que cubran diferentes combinaciones de oraciones, garantizando diversidad en las posibles soluciones iniciales.

La evaluación de la calidad de los resúmenes generados se lleva a cabo mediante una función de aptitud que considera dos características fundamentales: la frecuencia de palabras relevantes y la posición de las oraciones. Estas características son clave para identificar las oraciones más representativas del texto, ya que las palabras frecuentes tienden a reflejar los temas principales del documento y las oraciones en posiciones estratégicas, como las iniciales o finales, suelen contener información esencial (Matías-Mendoza et al., 2020). Para seleccionar los cromosomas que pasarán a la siguiente generación, se utiliza el método de selección por ruleta. Este enfoque asigna probabilidades de selección proporcionales a la calidad de cada cromosoma, de modo que las soluciones más prometedoras tienen mayores oportunidades de participar en los procesos de cruce y mutación.

El cruce, diseñado específicamente para la generación de resúmenes automáticos, combina los genes de dos cromosomas padres. Solo se seleccionan los genes con valor 1, lo que significa que se incluyen únicamente las oraciones marcadas como relevantes. Cuando ambos padres tienen un gen con valor 1 en la misma posición, este tiene una mayor probabilidad de ser heredado por el cromosoma hijo. Además, se controla que los resúmenes generados cumplan con el número mínimo de palabras requerido, lo cual, en el caso del corpus TeMário, corresponde al 30% del texto original.

La mutación se realiza mediante un enfoque de doble inversión. En la primera etapa, se alteran los genes con valor 1 para incluir o excluir ciertas oraciones. Posteriormente, se invierten los genes con valor 0 para añadir nuevas oraciones potencialmente relevantes o ajustar el contenido del resumen. Tras cada mutación, se verifica que el resumen cumpla con el tamaño mínimo requerido. Si no es así, el proceso continúa hasta que se alcance el número de palabras necesario.

A continuación, se describe la función de aptitud utilizada, al cual está dividida en frecuencia de las palabras más relevantes y en la posición de la oración.

La idea de esta fórmula de la posición de las oraciones se basa en que, si todas las oraciones tuvieran la misma importancia, su valor se representaría como una línea recta que indica uniformidad en la relevancia. Utilizando el punto medio de esta línea, se puede calcular la pendiente para ajustar la importancia de las oraciones de manera gradual. Esto permite determinar la relevancia relativa de una oración en comparación con las siguientes. De esta forma, la pendiente refleja si se otorga mayor importancia a las primeras o a las últimas oraciones: una pendiente negativa implica que las primeras oraciones son más importantes (por ejemplo, una pendiente de -1 indica una inclinación hacia la derecha a 45 grados). Una pendiente de cero significa que todas las oraciones tienen la misma importancia, mientras que una pendiente positiva señala que las últimas oraciones son más relevantes (una pendiente de 1 implica una inclinación hacia la derecha a 45 grados). En la tabla 2 se mientras las fórmulas utilizadas en la función de aptitud.

Tabla 2. Parámetros de la función de aptitud

Frecuencia de las palabras	Posición de las oraciones
$\beta = \frac{\sum_{p=\{word \in S\}}^{m} frecuency(p,T)}{\sum_{q=\{word \in T\}}^{m} frequency(q,T)}$	$\delta = \frac{\sum_{\lvert c_i \rvert=1}^{n} m(i-x)+x}{\sum_{j=1}^{k} m(j-x)+1}$, $x = 1 + \frac{(n-1)}{2}$
Donde S es el resumen a generar, m umbral máximo de palabras de un resumen, T las palabras más relevantes del texto original, por un lado se consideran la frecuency (w,T), para determinar la frecuancia, y la expresividad se representa si sólo se consideran las diferentes palabras que puede tener el resumen $\{word \in S\}$.	Para suavisar la posición de las oraciones se utiliza la ecución lineal con pendiente m (a la cual le daremos diferentes valores que van desde -0.5 a -1). Para un texto con n oraciones, si la oración i es seleccionada para el resumen, entonces su relevancia se define como $m(i-x)+x$, donde $= 1 + \frac{(n-1)}{2}$ y m es la pendiente por descubrir. Con el fin de normalizar la medida de la posición de la oración , se calcula la importancia de las primeras oraciones, donde es el número de oraciones seleccionadas

3.4 Experimentación

Cuando se utiliza un método de optimización, es fundamental realizar múltiples experimentos para garantizar la validez de los resultados, ya que este tipo de técnicas trabaja con soluciones aproximadas que dependen de varios factores, como los parámetros iniciales, las condiciones del problema y las características del conjunto de datos. Por esta razón, se llevaron a cabo dos experimentos para cada prueba. Por lo que los resultados mostrados en las tablas y gráficas son el promedio de los dos experimentos realizados.

La función de aptitud utilizada considera la importancia de la posición de las oraciones, para ello se utiliza la pendiente de la recta, por lo que es un parámetro por considerar al momento de hacer pruebas, en este trabajo se consideran las siguientes: m=-0.25, m=-0.3, m=-0.45, m=-0.5, m=-0.6, m=-0.65, m= -0.7, m=-0.75, m=-0.8, m= -0.85, m=-0.9, m=-0.95, m=-1.

La herramienta de evaluación utilizada en este trabajo fue ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*), un conjunto de métricas ampliamente usadas para evaluar la calidad de resúmenes automáticos en tareas de PLN (Lin, 2004).

ROUGE compara el resumen generado automáticamente con uno o varios resúmenes de referencia, hechos por humanos, midiendo la superposición de palabras, frases o secuencias entre ambos. En este caso, el corpus TeMário incluye un resumen humano que se usará como referencia para evaluar los resultados del resumen generado por el método propuesto.

Las principales métricas evaluadas son:

- Precisión: Proporción de las palabras relevantes seleccionadas en el resumen generado en relación con las palabras totales en ese resumen.
- Recuerdo: Proporción de palabras clave del resumen de referencia que están presentes en el resumen generado.
- *F-measure*: Combina precisión y recuerdo, proporcionando una métrica equilibrada que refleja tanto la exactitud como la exhaustividad del resumen.

ROUGE mide qué tan bien el resumen generado coincide con el de referencia, evaluando tanto la precisión como el recuerdo, y combinándolos en el *F-measure.* Para realizar esta evaluación, ROUGE utiliza fórmulas que calculan la coincidencia de palabras o secuencias (n-gramas) presentes en ambos resúmenes, proporcionando una evaluación objetiva y detallada de la calidad del resumen generado.

Las métricas más comunes en ROUGE incluyen:

1. ROUGE-N: Se basa en la coincidencia de n-gramas (secuencias de n palabras consecutivas) entre el resumen generado y el de referencia.
2. ROUGE-1 mide coincidencias de palabras individuales (unigramas).
3. ROUGE-2 mide coincidencias de pares de palabras consecutivas (bi-gramas).
4. ROUGE-N puede extenderse para n-gramas de mayor tamaño.

En este trabajo, los resultados se presentan utilizando la métrica ROUGE-1 y la métrica *F-measure* para evaluar la calidad de los resúmenes generados.

4. RESULTADOS

En esta sección se presentan los resultados obtenidos al aplicar el método propuesto para la generación automática de resúmenes en portugués. Los resultados se analizan en función de la calidad de los resúmenes generados, comparados con los resúmenes de referencia a través de métricas de evaluación ROUGE. Además, se muestran los efectos de las diferentes configuraciones del método en aspectos clave como el modelo de texto, preprocesamiento y la tokenización.

4.1 Experimento 1

El objetivo de este experimento es evaluar la calidad de la generación de resúmenes automáticos a partir del corpus TeMário, aplicando técnicas de pre-procesamiento y tokenización por oraciones. La Tabla 3 muestra los resultados obtenidos, de los cuales se observa que el número de n-gramas no influye significativamente en los resultados, ya que el modelo de palabras alcanza los mejores resultados con 0.4575 en *f-measure*. En cuanto a la relevancia de las oraciones, se confirma que las primeras oraciones son las más relevantes, dado que los resultados óptimos para el modelo de palabras se obtienen en un rango de pendientes entre -0.7 y -1.

Tabla 3. Resultados por modelo de texto en relación con la importancia de las oraciones

Mod Tex/m=	-0.25	-0.3	-0.45	-0.5	-0.55	-0.6	-0.65	-0.7	-0.75	-0.8	-0.85	-0.9	-0.95	-1
Palabras	0.4482	0.4535	0.4527	0.4530	0.4546	0.4513	0.4541	0.4567	0.4562	0.4543	0.4549	0.4525	0.4505	**0.4575**
Bi-gramas	0.4510	0.4483	0.4551	0.4512	0.4364	0.4518	0.4563	**0.4566**	0.4532	0.4540	0.4534	0.4546	0.4541	0.4553
Tri-gramas	0.4491	0.4460	0.4486	0.4491	0.4479	0.4492	0.4504	0.4473	0.4471	0.4464	0.4476	**0.4507**	0.4496	0.4484
Cuatri-gramas	0.4469	0.4472	0.4504	0.4509	0.4500	0.4497	0.4486	0.4488	0.4517	0.4499	0.4528	**0.4533**	0.4515	0.4502

4.2 Experimento 2

El objetivo de este experimento es evaluar la calidad de los resúmenes generados a partir del corpus TeMário sin pre-procesamiento y tokenización por oraciones. La Tabla 4 muestra los resultados obtenidos, al igual que los resultados del experimento anterior se observa que el número de n-gramas no influye significativamente en el rendimiento, ya que el modelo de palabras individuales alcanza los mejores resultados con 0.4575 en *f-measure*. En cuanto a la relevancia de las oraciones, cuando no se realiza preprocesamiento el método tiende a elegir oraciones que no se encuentran al inicio del texto. Sin embargo, para el modelo palabras se confirma que las primeras oraciones son las más relevantes, dado que los resultados óptimos para el modelo de palabras se obtienen con una pendiente de -0.95.

Tabla 4. Resultados por modelo de texto en relación con la importancia de las oraciones

Mod Tex/m=	-0.25	-0.3	-0.45	-0.5	-0.55	-0.6	-0.65	-0.7	-0.75	-0.8	-0.85	-0.9	-0.95	-1
Palabras	0.4470	0.4515	0.4496	0.4495	0.4489	0.4500	0.4490	0.4507	0.4501	0.4564	0.4554	0.4544	**0.4575**	0.4467
Bi-gramas	0.4508	0.4514	0.4529	0.4536	**0.4569**	0.4498	0.4528	0.4508	0.4491	0.4529	0.4548	0.4545	0.4527	0.4543
Tri-gramas	0.4483	0.4485	**0.4525**	0.4506	0.4504	0.4494	0.4487	0.4518	0.4486	0.4500	0.4505	0.4495	0.4490	0.4484
Cuatri-gramas	0.4446	0.4462	0.4452	0.4476	0.4491	0.4476	0.4500	0.4464	0.4477	0.4470	0.4487	**0.4518**	0.4492	0.4500

4.3 Experimento 3

Este experimento busca evaluar la calidad de los resúmenes automáticos generados a partir del corpus TeMário, sin aplicar pre-procesamiento y utilizando tokenización por párrafos. A diferencia de la tokenización por oraciones, la tokenización por párrafos permite capturar ideas más completas y analizar la coherencia y continuidad temática dentro de un bloque mayor de texto, lo cual es especialmente útil para resúmenes donde se requiere una visión global y contextualizada de la información.

El experimento mantiene los mismos parámetros en cuanto a modelo de texto y pendiente utilizados en experimentos previos, para asegurar la comparabilidad de resultados. La Tabla 5 muestra los resultados obtenidos, para el uso de párrafos el mejor resultado se obtiene con el modelo de texto bi-gramas con 0.4686 (*f-measure*) en una pendiente de -0.65.

Tabla 5. Resultados por modelo de texto en relación con la importancia de las oraciones

Mod Tex/m=	-0.25	-0.3	-0.45	-0.5	-0.55	-0.6	-0.65	-0.7	-0.75	-0.8	-0.85	-0.9	-0.95	1
Palabras	0.4438	0.4439	0.4499	0.4473	0.4426	0.4469	0.4488	0.4479	0.4496	0.4462	0.4453	0.4480	0.4477	**0.4517**
Bi-gramas	0.4445	0.4441	0.4558	0.4529	0.4571	0.4538	**0.4586**	0.4523	0.4542	0.4502	0.4502	0.4558	0.4540	0.4527
Tri-gramas	0.4476	0.4488	0.4515	0.4490	0.4479	0.4484	0.4476	0.4496	0.4471	**0.4560**	0.4465	0.4481	0.4472	0.4486
Cuatri-gramas	0.4451	0.4447	0.4485	0.4480	0.4503	0.4486	**0.4514**	0.4446	0.4490	0.4486	0.4472	0.4503	0.4488	0.4497

Los experimentos realizados con el corpus TeMário proporcionan una comprensión detallada de los efectos del pre-procesamiento, la tokenización (por oraciones y párrafos), y los modelos de texto en la calidad de los resúmenes automáticos en portugués.

En primer lugar, el pre-procesamiento no demostró ser un factor decisivo en la mejora de la calidad de los resúmenes. Aunque su ausencia provocó una selección menos estructurada de oraciones, los resultados sin pre-procesamiento fueron comparables en precisión y recuerdo a aquellos con pre-procesamiento. Esto indica que, si bien el pre-procesamiento ayuda a estructurar el resumen, su impacto en los resultados generales es limitado.

La tokenización mostró una influencia importante en la coherencia y la relevancia del resumen. La tokenización por oraciones se desempeñó bien en los primeros experimentos, con un rendimiento óptimo alcanzado seleccionando las primeras oraciones del texto, reflejando la estructura típica de los textos periodísticos, donde las primeras oraciones suelen concentrar información relevante. Sin embargo, la tokenización por párrafos en el tercer experimento resultó ser aún más eficaz para capturar ideas completas y cohesivas, alcanzando el mejor puntaje de todos los experimentos (0.4686 en *f-measure*) con el modelo de bi-gramas. Esto sugiere que, en contextos donde se requiere una comprensión más global del contenido, la tokenización por párrafos ofrece ventajas significativas.

En cuanto a los modelos de texto, el modelo de palabras individuales mostró buenos resultados en la mayoría de las configuraciones, pero el modelo de bi-gramas en tokenización por párrafos superó en desempeño a los otros, permitiendo mantener la coherencia semántica a través de relaciones contextuales entre palabras.

5. CONCLUSIÓN

En este trabajo se plantea el problema de la escasez de investigaciones y herramientas especializadas para la generación automática de resúmenes en portugués, un lenguaje hablado por millones de personas en diversos países. Este vacío en la investigación limita el acceso a soluciones de PLN adaptadas a las características y necesidades de este lenguaje. Por tanto, este trabajo busca abordar esa brecha, proporcionando un análisis de parámetros óptimos y métodos que mejoran la calidad de los resúmenes generados en portugués.

Los resultados obtenidos destacan la importancia de seleccionar adecuadamente los parámetros de tokenización y modelo de texto para maximizar la precisión y relevancia en los resúmenes generados. Los resultados muestran que para oraciones el modelo bolsa de palabras es el más adecuado. Además, el análisis de la pendiente óptima reveló que un rango entre -0.65 y -0.95 permite identificar con mayor precisión las oraciones relevantes, lo cual es un avance relevante para el portugués, un lenguaje que tradicionalmente ha sido menos estudiado en el área de PLN en comparación con el inglés.

Por otra parte, aunque el preprocesamiento contribuye a estructurar el texto, los experimentos revelaron que su ausencia no impacta de manera significativa en la calidad del resumen final. Esto indica una notable flexibilidad en el manejo de textos en portugués, permitiendo simplificar el proceso sin comprometer los resultados. De hecho, los resúmenes generados a partir de textos sin preprocesamiento mostraron un rendimiento ligeramente superior en comparación con aquellos preprocesados. Este trabajo evidencia que, al comprender y ajustar los parámetros específicos para el lenguaje portugués, es posible mejorar la precisión y

relevancia de los resúmenes generados automáticamente, proporcionando una base sólida para el desarrollo futuro de métodos para la generación automática de resúmenes que respondan mejor a las necesidades de los hablantes de portugués.

REFERENCIAS

Albuquerque, A., (2010). *El valor económico del portugués: lengua de conocimiento con influencia global (ARI).* Real Inst. Elcano.

Amancio, D.R., Nunes, M.G., Oliveira Jr, O.N., Costa, L. da F., (2012). Extractive summarization using complex networks and syntactic dependency. *Phys. Stat. Mech*. Its Appl. 391, 1855–1864.

Antiqueira, L., (2007). *Desenvolvimento de técnicas baseadas em redes complexas para sumarização extrativa de textos.* Universidade de São Paulo.

Asevedo Nóbrega, F.A., Salgueiro Pardo, T.A., (2017). Update Summarization for Portuguese, in: 2017 Brazilian Conference on Intelligent Systems (BRACIS). *Presented at the 2017 Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 348–353. https://doi.org/10.1109/BRACIS.2017.49

Cabral, L. de S., Lins, R.D., Mello, R.F., Freitas, F., Ávila, B., Simske, S., Riss, M., (2014). A platform for language independent summarization, in: *Proceedings of the 2014 ACM Symposium on Document Engineering. ACM*, pp. 203–206.

Califano, B., (2015). Los medios de comunicación, las noticias y su influencia sobre el sistema político. *Rev. Mex. Opinión Pública* 19, 61–79. https://doi.org/10.1016/j.rmop.2015.02.001

Cardoso, P.C., Maziero, E.G., Jorge, M.L., Seno, E.M., Di Felippo, A., Rino, L.H., Nunes, M.G., Pardo, T.A., (2011). CSTnews-a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese, *in: Proceedings of the 3rd RST Brazilian Meeting*. pp. 88–105.

Cavalieri, D.C., Bastos-Filho, T., Palazuelos-Cagigas, S.E., Sarcinelli-Filho, M., (2015). On combining language models to improve a text-based human-machine interface. *Int. J. Adv. Robot*. Syst. 12, 170.

de Oliveira, M.A., Guelpeli, M.V., (2011). *BLMSumm–Métodos de Busca Local e Metaheurísticas na Sumarização de Textos.*

El-Kassas, W.S., Salama, C.R., Rafea, A.A., Mohamed, H.K., (2021). Automatic text summarization: A comprehensive survey. *Expert Syst.* Appl. 165, 113679.

Garcia, G.L., Paiola, P.H., Jodas, D.S., Sugi, L.A., Papa, J.P., (2024). Text Summarization and Temporal Learning Models Applied to Portuguese Fake News Detection in a Novel Brazilian Corpus Dataset, in: Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H.G., Amaro, R. (Eds.), *Proceedings of the 16th International Conference on Computational Processing of Portuguese* - Vol. 1. Presented at the PROPOR 2024, Association for Computational Linguistics, Santiago de Compostela, Galicia/Spain, pp. 86–96.

Isonuma, M., Fujino, T., Mori, J., Matsuo, Y., Sakata, I., (2017). Extractive Summarization Using Multi-Task Learning with Document Classification, in: Palmer, M., Hwa, R., Riedel, S. (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Presented at the EMNLP 2017, Association for Computational Linguistics, Copenhagen, Denmark, pp. 2101–2110. https://doi.org/10.18653/v1/D17-1223

Jin, H., Zhang, Y., Meng, D., Wang, J., Tan, J., (2024). *A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods*. https://doi.org/10.48550/arXiv.2403.02901

Lin, C.-Y., (2004). Rouge: A package for automatic evaluation of summaries. Presented at the Text summarization branches out: *Proceedings of the ACL-04 workshop*, Barcelona, Spain.

Lins, A.A., Carvalho, C.S., Bomfim, F.D.C.J., de Carvalho Bentes, D., Pinheiro, V., (2024). CLSJUR.BR - A Model for Abstractive Summarization of Legal Documents in Portuguese Language based on Contrastive Learning, *Proceedings of the 16th International Conference on Computational Processing of Portuguese* - Vol. 1. Presented at the PROPOR 2024, Association for Computational Lingustics, Santiago de Compostela, Galicia/Spain, pp. 321–331.

Margarido, P.R., Pardo, T.A., Antonio, G.M., Fuentes, V.B., Aires, R., Aluísio, S.M., Fortes, R.P., (2008). Automatic summarization for text simplification: Evaluating text understanding by poor readers, in: Companion *Proceedings of the XIV Brazilian Symposium on Multimedia and the Web. ACM*, pp. 310–315.

Matías-Mendoza, G.A., Ledeneva, Y., García-Hernández, R. A., (2020). Determining the importance of sentence position for automatic text summarization. *J. Intell. Fuzzy Syst.* 39, 2421–2431. https://doi.org/10.3233/JIFS-179902

Maziero, E.G., Uzêda, V.R., Pardo, T.A.S., Nunes, M. das G.V., (2007). TeMário 2006: Estendendo o Córpus TeMário. *Série de Relatórios do NILC*. NILC-TR-07-06. São Carlos-SP, Agosto, 8p.

Mihalcea, R., Tarau, P., (2005). A Language Independent Algorithm for Single and Multiple Document Summarization, in: *Companion Volume to the Proceedings of Conference Including Posters/Demos and Tutorial Abstracts*. Presented at the IJCNLP 2005.

Mussandi, J., Wichert, A., (2024). NLP Tools for African Languages, P*roceedings of the 16th International Conference on Computational Processing of Portuguese* - Vol. 2. Presented at the PROPOR 2024, Association for Computational Lingustics, Santiago de Compostela, Galicia/Spain, pp. 73–82.

Neri-Mendoza, V., Ledeneva, Y., García-Hernández, R.A., Hernández-Castañeda, Á., (2024). Multi-document Text Summarization through Features Relevance Calculation. *Comput. Sist*. 28. https://doi.org/10.13053/cys-28-3-5201

Nunes, M. das G.V., Aluísio, S.M., Pardo, T.A.S., 2010. Um panorama do Núcleo Interinstitucional de Linguística Computacional às vésperas de sua maioridade. *Linguamática* 2, 13–27.

Oliveira, M.A., Guelpeli, M.V.C., (2012). The performance of BLMSumm: Distinct languages with antagonistic domains and varied compressions, in: *2012 IEEE International Conference on Information Science and Technology. Presented at the 2012 IEEE International Conference on Information Science and Technology*, pp. 609–614. https://doi.org/10.1109/ICIST.2012.6221717

Orrú, T., Rosa, J.L.G., de Andrade Netto, M.L., (2006). SABIO: an automatic portuguese text summarizer through artificial neural networks in a more biologically plausible model, in: *International Workshop on Computational Processing of the Portuguese Language*. Springer, pp. 11–20.

Paiola, P.H., de Rosa, G.H., Papa, J.P., (2022). Deep Learning-Based Abstractive Summarization for Brazilian Portuguese Texts, in: Xavier-Junior, J.C., Rios, R.A. (Eds.), *Intelligent Systems*. Springer International Publishing, Cham, pp. 479–493. https://doi.org/10.1007/978-3-031-21689-3_34

Pardo, T.A.S., Antiqueira, L., Nunes, M. das G.V., Oliveira, O.N., da Fontoura Costa, L., (2006). Modeling and Evaluating Summaries Using Complex Networks, *Computational Processing of the Portuguese Language. Springer*, Berlin, Heidelberg, pp. 1–10. https://doi.org/10.1007/11751984_1

Pardo, T.A.S., Rino, L.H.M., (2003). *TeMário: Um corpus para sumarização automática de textos. São Carlos Universidade São Carlos Relatório Téc*.

Rino, L.H.M., Pardo, T.A.S., Nascimento Silla, C., Kaestner, C.A.A., Pombo, M., (2004). A Comparison of Automatic Summarizers of Texts in Brazilian Portuguese, *Advances in Artificial Intelligence – SBIA 2004. Springer*, Berlin, Heidelberg, pp. 235–244. https://doi.org/10.1007/978-3-540-28645-5_24

Silva, R.R. da, Pardo, T.A.S., (2022). Building Contrastive Summaries of Subjective Text Via Opinion Ranking. Rev. *Informática Teórica E Apl*. 29, 11–34. https://doi.org/10.22456/2175-2745.118372

Supriyono, Wibawa, A.P., Suyono, Kurniawan, F., (2024). A survey of text summarization: Techniques, evaluation and challenges. *Nat. Lang. Process.* J. 7, 100070. https://doi.org/10.1016/j.nlp.2024.100070