TecnoLógicas
ISSN: 0123-7799
ISSN: 2256-5337
tecnologicas@itm.edu.co
Instituto Tecnológico Metropolitano
Colombia

Comparación de algoritmos de resumen de texto para el procesamiento de editoriales y noticias en español

López-Trujillo, Sebastián; Torres-Madroñero, Maria C.
Comparación de algoritmos de resumen de texto para el procesamiento de editoriales y noticias en español
TecnoLógicas, vol. 24, núm. 51, e1816, 2021
Instituto Tecnológico Metropolitano, Colombia
Disponible en: https://www.redalyc.org/articulo.oa?id=344265925012
DOI: https://doi.org/10.22430/22565337.1816

Artículos de investigación

Comparación de algoritmos de resumen de texto para el procesamiento de editoriales y noticias en español

Comparison of Text Summarization Algorithms for Processing Editorials and News in Spanish

Sebastián López-Trujillo
Instituto Tecnológico Metropolitano, Colombia
Sebastianlopez249178@correo.itm.edu.co
https://orcid.org/0000-0003-1708-2834

Maria C. Torres-Madroñero
Instituto Tecnológico Metropolitano, Colombia
mariatorres@itm.edu.co
https://orcid.org/0000-0002-9795-2459

DOI: https://doi.org/10.22430/22565337.1816
Redalyc: https://www.redalyc.org/articulo.oa?id=344265925012

Recepción: 09 Diciembre 2020
Aprobación: 20 Mayo 2021
Publicación: 10 Junio 2021

Resumen:

El lenguaje se ve afectado, no solo por las reglas gramaticales, sino también por el contexto y las diversidades socioculturales, por lo cual, el resumen automático de textos (un área de interés en el procesamiento de lenguaje natural - PLN), enfrenta desafíos como la identificación de fragmentos importantes según el contexto y el tipo de texto analizado. Trabajos anteriores describen diferentes métodos de resúmenes automáticos, sin embargo, no existen estudios sobre su efectividad en contextos específicos y tampoco en textos en español. En este artículo se presenta la comparación de tres algoritmos de resumen automático usando noticias y editoriales en español. Los tres algoritmos son métodos extractivos que buscan estimar la importancia de una frase o palabra a partir de métricas de similitud o frecuencia de palabras. Para esto se construyó una base de datos de documentos donde se incluyeron 33 editoriales y 27 noticias, obteniéndose un resumen manual para cada texto. La comparación de los algoritmos se realizó cuantitativamente, empleando la métrica Recall-Oriented Understudy for Gisting Evaluation. Asimismo, se analizó el potencial de los algoritmos seleccionados para identificar los componentes principales del texto. En el caso de las editoriales, el resumen automático debía incluir un problema y la opinión del autor, mientras que, en las noticias, el resumen debía describir las características temporales y espaciales de un suceso. En términos de porcentaje de reducción de palabras y precisión, el método que permite obtener los mejores resultados, tanto para noticias como para editoriales, es el basado en la matriz de similitud. Este método permite reducir en un 70 % los textos, tanto editoriales como noticiosos. No obstante, es necesario incluir la semántica y el contexto en los algoritmos para mejorar su desempeño en cuanto a precisión y sensibilidad.

Palabras clave: Procesamiento de lenguaje natural, *Recall-Oriented Understudy for Gisting Evaluation* , análisis de textos, minería de textos, resumen automático.

Abstract:

Language is affected not only by grammatical rules but also by the context and socio-cultural differences. Therefore, automatic text summarization, an area of interest in natural language processing (NLP), faces challenges such as identifying essential fragments according to the context and establishing the type of text under analysis. Previous literature has described several automatic summarization methods; however, no studies so far have examined their effectiveness in specific contexts and Spanish texts. In this paper, we compare three automatic summarization algorithms using news articles and editorials in Spanish. The three algorithms are extractive methods that estimate the importance of a phrase or word based on similarity or word frequency metrics. A document database was built with 33 editorials and 27 news articles, and three summaries of each text were manually extracted employing the three algorithms. The algorithms were quantitatively compared using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric. We analyzed the algorithms' potential to identify the main components of a text. In the case of editorials, the automatic summary should include a problem and the author's opinion. Regarding news articles, the summary should describe the temporal and spatial characteristics of an event. In terms of word reduction percentage and accuracy, the method based on the similarity matrix produced the best results and can achieve a 70 % reduction in both cases (i.e., news and

editorials). However, semantics and context should be incorporated into the algorithms to improve their performance in terms of accuracy and sensitivity.

Keywords: Natural language processing, Recall-Oriented Understudy for Gisting Evaluation, Text Analysis, Text Mining, Automatic Summarization.

Highlights

- Existen pocos estudios de resumen automático de textos en español.
- Se comparan algoritmos extractivos de resumen en noticias y editoriales.
- Las métricas de similitud permiten reducir en un 70 % los textos.

1. INTRODUCCIÓN

El procesamiento de lenguaje natural (PLN) es un área de conocimiento dentro del campo de la inteligencia artificial [1]. El PLN tiene su origen en la búsqueda de soluciones para la interpretación de los diferentes idiomas del mundo, búsqueda que llevó al desarrollo de los traductores. Posteriormente, surgieron aplicaciones de PLN con el propósito de mejorar la comunicación con las máquinas. Entre estas aplicaciones se encuentran los sistemas de análisis de textos [1]. Dada la diversidad de léxicos, reglas gramaticales e intenciones de la comunicación escrita, existen diferentes retos a los cuales se deben enfrentar los algoritmos de PLN para el análisis de textos, tales como la ambigüedad, la interpretación de expresiones callejeras y muletillas, entre otros [2]-[4]. Uno de los enfoques o etapas de PLN para análisis de textos es la construcción de resúmenes: métodos que buscan extraer de forma automática la información más relevante de un texto para ser presentada a un usuario [5].

Los enfoques para el resumen automático de textos se pueden clasificar en cuatro categorías: métodos basados en términos, métodos basados en frases, métodos de taxonomía de patrones y métodos basado en conceptos. Los métodos basados en términos realizan la identificación de palabras que tienen un significado semántico. Este enfoque no tiene ninguna restricción del contexto y presenta dificultades con las palabras que son polisémicas – palabras cuyo significado depende del contexto [2] - y las palabras sinónimas [6], [7]. Por su parte, el enfoque basado en frases permite que la identificación de palabras sea menos ambigua, ya que estas frases se pueden considerar como información. La dificultad de este método está en la identificación del tema principal de un texto [7]. Los métodos basados en taxonomía de patrones analizan los documentos a partir de la cantidad de veces que se repite una palabra o una frase. Sin embargo, esos métodos sufren de incongruencias ya que no todas las palabras o frases que se repitan son relevantes [7], [8]. Finalmente, los métodos basados en conceptos analizan de forma estadística frases y palabras para identificar la relevancia.

En estos métodos, primero se analiza la estructura semántica de la oración, luego se construye un árbol gráfico de conceptos, a partir del cual se extraen los conceptos que permiten resumir el texto [7], [9].

Este artículo se enfoca en los métodos basados en taxonomía de patrones, dada la diversidad de técnicas encontradas en la literatura, así como su desempeño. En estos métodos se pueden identificar dos fases: extracción y abstracción [10]. En la fase de extracción se asignan etiquetas o puntajes a cada palabra del texto. El puntaje depende de la cantidad de veces que se repita la palabra o de la posición en que se encuentre en el texto. En esta etapa se separan los conectores con el fin de extraer únicamente las palabras que sean relevantes para entender el tema principal. La fase de abstracción emplea estadísticas y bases de datos entrenadas para identificar el tema principal y construir el resumen [9], [10]. Entre mejor esté entrenado el modelo o alimentada la base de datos se podrá tener una mejor interpretación del texto [5].

Reconociendo las posibilidades que tienen los sistemas de análisis de textos para diferentes aplicaciones, este artículo se centra en la comparación de algoritmos de resumen automático en el contexto de análisis

de noticias y editoriales en español. El análisis de texto automático puede ser incorporado a sistemas e-learning para la calificación automática de textos largos [11], así como apoyo a las tareas de análisis de datos en metodologías de investigación cualitativas [12]. En la literatura se encuentran diferentes métodos de resumen automático, evaluados principalmente con bases de datos en inglés [13]. Este articulo realiza una comparación de métodos de resumen automático en contextos específicos, como son el análisis de noticias y editoriales, particularmente en español. A diferencia del inglés, la comunicación escrita en español se caracteriza por su verbosidad, utilizada para resaltar, modificar o rellenar el argumento principal de un texto, lo cual impone retos para los sistemas automáticos de resumen de los mismos. Ejemplos de investigaciones orientados al uso de PLN en textos en español incluyen la identificación de errores léxico-sintácticos [14] y el resumen de noticias [13]. A diferencia de la comparación de algoritmos presentada en [13], este estudio se centra en la comparación de resúmenes en dos tipos de documento: editoriales y noticias. A diferencia de las noticias, las editoriales se estructuran desde la opinión del escritor sobre una problemática específica, desarrollando los argumentos que soportan su punto de vista. Sin embargo, y de acuerdo con la literatura, los métodos basados en taxonomía de patrones son aplicables a diferentes contextos [10].

Para la comparación presentada en este artículo, se construyó una base de datos de noticias y editoriales seleccionadas de forma aleatoria sin una temática predefinida, desde revista, magazines y periódicos digitales en español. Además, se elaboraron resúmenes de referencia para cada texto considerando las características particulares de cada tipo de documento. En el caso de editoriales, al tratarse de artículos de opinión, se busca identificar un problema o situación, el punto de vista del autor y los argumentos que sustentan su punto de vista. Para las noticias se espera identificar un suceso o hecho, con la identificación temporal y espacial, así como los protagonistas. El objetivo de este artículo es comparar tres algoritmos basados en taxonomía de patrones empleando el porcentaje de reducción de palabras y una métrica de evaluación de resúmenes de texto ampliamente usada en la literatura denominada ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [7].

En la siguiente sección se describen en detalle los principios teóricos de los tres algoritmos seleccionados. En la sección de experimentos se presenta la construcción de la base de datos, detalles de la implementación de los algoritmos en Python y la descripción de las métricas de evaluación. Posteriormente, se presentan los resultados obtenidos tanto para noticias como para las editoriales. Finalmente, se presentan las conclusiones y posibles trabajos futuros de esta investigación.

2. METODOLOGÍA

Los algoritmos comparados en este artículo se basan en la extracción de palabras, eliminando conectores de los documentos para posteriormente asignar un peso a cada frase, determinado a partir de la frecuencia o la similitud de las palabras. Esta sección describe los algoritmos empleados en este estudio comparativo.

2.1 Algoritmo basado en matriz de similitud

Dentro de las diferentes técnicas de resumen automático de textos, usualmente se emplean enfoques extractivos, los cuales seleccionan las frases que representan las ideas principales, eliminando, inicialmente, los conectores y, posteriormente, determinando la importancia o peso de la frase. Entre estos métodos se encuentran los basados en métricas de similitud [15], [16]. Entre las distancias empleadas para calcular la similitud de las oraciones se encuentra la distancia coseno [17]. Esta matriz es empleada por el algoritmo PageRank [15], [16] para calcular los pesos de las frases. Finalmente, se seleccionan las N frases con el peso más alto para construir el resumen. El método se describe en la Figura 1.

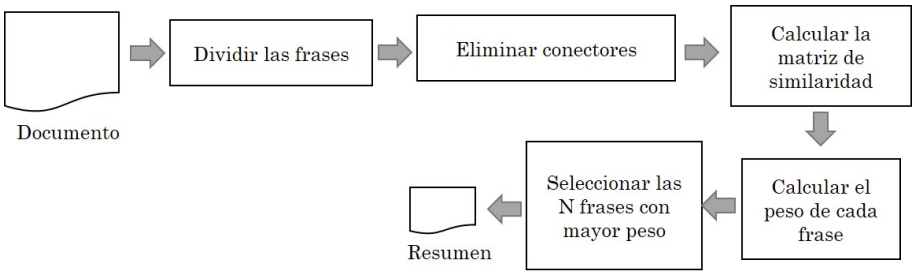

FIGURA 1.
Algoritmo de resumen automático basado en matriz de similitud.
Fuente: elaboración propia.

2.2 Algoritmos basados en frecuencia de las palabras

Los algoritmos basados en frecuencia calculan un puntaje para cada frase de acuerdo con la frecuencia de las palabras (TF) que la componen. Para un algoritmo, estos puntajes son calculados al utilizar las técnicas de frecuencia de término – frecuencia inversa de documento (TF-IDF, por sus siglas en inglés) [18], que se basa en la frecuencia de una palabra en el documento y en un conjunto de documentos. Para calcular el puntaje TF-IDF, primero se debe hacer una eliminación de conectores, ya que son los que más se repiten, y determinar el número de veces *nx* que un término *x* aparece en un documento y se calcula el término TF dado por (1), donde *n* es el número total de términos en el documento. Adicionalmente, se debe calcular el IDF de acuerdo con (2), donde m es el número total de documentos considerados y *mx* es el número de documentos que incluyen el término *x*. Finalmente, ambos índices se combinan mediante (3). Para el cálculo de TF-IDF:

$$TF = \frac{nx}{n} \tag{1}$$

$$IDF = log_e \frac{m}{mx} \tag{2}$$

$$TFIDF = TF * IDF \tag{3}$$

En este artículo se implementó también un algoritmo extractivo basado en el índice *TF* de (1). La Figura 2 describe los métodos basados en frecuencia de palabras para el resumen automático de textos.

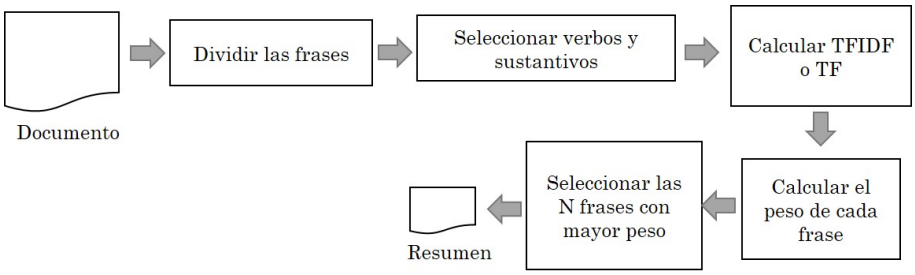

FIGURA 2.
Algoritmos de resumen automático basado en frecuencia de palabras.
Fuente: elaboración propia.

2.3 Base de datos

Para el desarrollo de los experimentos se construyó una base de datos a partir de editoriales y noticias en español sin ninguna temática especifica. Las editoriales fueron extraídas de revistas científicas, revistas de opinión y periódicos digitales, siendo seleccionadas, de forma aleatoria, 33 editoriales. Por su parte, las noticias se extrajeron de periódicos digitales, donde fueron seleccionados 27 textos al azar. Todos los documentos tienen más de 300 palabras. Para cada documento se construyó un resumen manual teniendo en cuenta las características particulares de cada tipo de documento. En el caso de las editoriales, se define que el resumen debe incluir al menos la descripción de un problema o asunto, así como el punto de vista del autor, con sus argumentos principales. Para las noticias, el resumen incluye un suceso, con la descripción de tiempo y lugar en el que ocurre, así como los protagonistas o actores del suceso.

2.4 Implementación de algoritmos

Los algoritmos comparados en este estudio son:

-Algoritmo de resumen automático basado en la matriz de similitud (Figura 1).
-Algoritmo de resumen automático basado en frecuencia de palabras calculado por el índice TF (Figura 2 – (1)).
-Algoritmo de resumen automático basado en frecuencia de palabras calculado por el índice TFIDF (Figura 2 – (3)).

Todos los algoritmos utilizaron la librería de Python NTLK (Natural Language Toolkit), la cual cuenta con las herramientas necesarias para realizar la extracción y/o abstracción de los textos. NLTK es un paquete de herramientas libres para trabajar el PLN [18]. Los tres algoritmos que se mencionan a continuación utilizan el NLTK para la fase de extracción donde se hace la división de palabras y frases de los documentos. Este también cuenta con un diccionario de palabras que son muy utilizadas o repetitivas en los lenguajes, estas palabras son los conectores y palabras gramaticales [14], las cuales deben ser eliminadas para tener un mejor concepto de las palabras importantes de los textos.

Para los tres algoritmos se emplearon implementaciones basadas en la librería NLTK [19]. En los tres casos, se empleó la función corpus, encargada de invocar el diccionario de palabras comunes en el lenguaje seleccionado, en este caso español. Para el caso del algoritmo basado en la matriz de similitud, se empleó la distancia coseno, calculada por la función *cluter.util*.

A partir de la matriz de similitud se asigna el peso a cada frase empleando el algoritmo PageRank (función pagerank de la librería networkx de Python) y se seleccionan las N frases con el peso más alto. Para estos

experimentos, N se selecciona igual a cinco. Para el caso de los algoritmos basados en la frecuencia de las palabras, TF-IDF y TF, se emplea la función pos_tag de la librería NLTK para clasificar las palabras en el documento, y seleccionar únicamente los verbos y los sustantivos. El TF-IDF se calcula de acuerdo con (3) y TF de acuerdo con (1).

2.5 Métricas de evaluación

Para la comparación de los tres algoritmos de resumen automático, basado en matriz de similitud, TF-IDF y TF, se emplearon dos métricas. Primero, se compararon los porcentajes de reducción de número de palabras. Adicionalmente, se calcula la métrica suplemento orientada al recuerdo para la evaluación de ROUGE, donde específicamente se emplearon la precisión, la sensibilidad y el valor F de ROUGE-1 y ROUGE-L [7]. ROUGE es un conjunto de métricas creadas para comparar resúmenes automáticos de textos con resúmenes de referencia creados por humanos, que se basa en los conceptos de precisión (P), sensibilidad (R, por su nombre en inglés Recall) y el valor F (F, por su nombre en inglés F-score) [7]. Por ejemplo, la métrica ROUGE-1 compara los textos con base a cada palabra; es decir, calcula P, S y F a partir de las palabras en común entre el resumen de referencia y el resumen automático. En cambio, ROUGE-L, se basa en la subsecuencia común más larga (LCS, por su nombre en inglés: Longest Common Subsequence) para el cálculo de las estadísticas [7]. Las ecuaciones (4), (5) y (6) resumen el cálculo de cada métrica [7], donde *nc* es el número de palabras (ROUGE-1) o LCS (ROUGE-L) en común entre el resumen de referencia y el resumen automático; *nra* es el número de palabras en el resumen automático; y *nrr* el número de palabras en el resumen de referencia. La precisión se interpreta como una medida de la relevancia de las palabras o LCS incluidas en el resumen automático. Por su parte, la sensibilidad cuantifica cuánto del resumen de referencia es capturado en el resumen automático. El valor F se calcula a partir de la precisión y la sensibilidad. El valor más alto posible de F es 1, representando una precisión y sensibilidad del 100 %, respectivamente.

$$P = \frac{nc}{nra} \tag{4}$$

$$R = \frac{nc}{nrr} \tag{5}$$

$$F = 2\frac{P * R}{P + R} \tag{6}$$

3. RESULTADOS Y DISCUSIÓN

En esta sección se presentan los resultados obtenidos a partir del resumen automático de las noticias y editoriales, empleando los tres algoritmos basados en matriz de similitud, TF-IDF y TF, y calculando el porcentaje de reducción de palabras y la métrica ROUGE. Se presentan los resultados, primero para las noticias, y luego para las editoriales.

3.1 Resultados para noticias

La Figura 3 presenta el porcentaje de reducción de palabras obtenidas en el resumen de las noticias empleando el método basado en matriz de similitud (SM), las métricas TF-IDF y TF, y el resumen de referencia (R). El algoritmo basado en la matriz de similitud presenta los porcentajes de reducción más altos en la mayoría de los documentos, incluso comparado con los porcentajes de reducción obtenidos por el resumen manual. En promedio, el algoritmo basado en la matriz de similitud obtuvo una reducción del 70 % de las palabras de las noticias, con una desviación estándar del 23 %. En el caso del resumen de referencia, el porcentaje de reducción promedio fue del 64 %. Para los métodos basados en frecuencia se obtuvo un 55 % y 56 % de reducción empleando TF-IDF y TF, respectivamente.

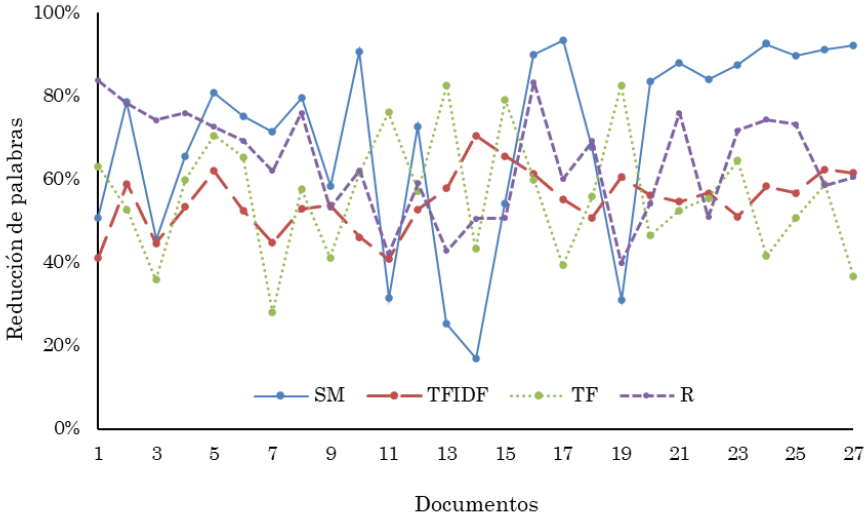

FIGURA 3.
Porcentaje de reducción de palabras en resúmenes de noticias empleando el método basado en matriz de similitud (SM), las métricas TF-IDF y TF, y el resumen de referencia (R)
Fuente: elaboración propia.

La Tabla 1 presenta el promedio (μ) y la desviación estándar (σ) para la precisión (P), sensibilidad (R) y valor F de la métrica ROUGE-1 para cada uno de los métodos evaluados. De acuerdo con la precisión, el mejor resumen fue obtenido por el método basado en la matriz de similitud con un 69 %. Sin embargo, en términos de sensibilidad, el mejor resultado lo presenta el algoritmo basado en TF con un 69 %. El valor F más alto es obtenido con el algoritmo TF-IDF con un 58 %. Al observar la desviación estándar de la Tabla 1, se ve que la precisión para todos los métodos es de 16 %. El valor más alto para R y F, lo tiene la matriz de similitud con un 24 % para R y un 15 % para F. ROUGE-L se presentan en la Tabla 2. En este caso, de acuerdo con la precisión, el método basado en la matriz de similitud como TF obtiene 43 % de precisión. El método TF obtiene el mejor resultado de acuerdo con la sensibilidad (54 %) y valor F (44 %). Las desviaciones estándar varían entre 13 % y 20 % para todas las métricas.

TABLA 1.
Promedio (μ) y desviación estándar (σ) para las métricas ROUGE-1 para los resúmenes automáticos de noticias empleando los algoritmos basados en similitud (SM), TF-IDF y TF.

	SM			TF-IDF			TF		
	P	R	F	P	R	F	P	R	F
μ	0.69	0.51	0.53	0.53	0.68	0.58	0.56	0.69	0.57
σ	0.16	0.24	0.15	0.16	0.15	0.12	0.16	0.21	0.12

Fuente: elaboración propia.

TABLA 2.
Promedio (μ) y desviación estándar (σ) para las métricas ROUGE-L para los resúmenes automáticos de noticias empleando los algoritmos basados en similitud (SM), TF-IDF y TF.

	SM			TF-IDF			TF		
	P	R	F	P	R	F	P	R	F
μ	0.43	0.33	0.33	0.39	0.50	0.42	0.43	0.54	0.44
σ	0.13	0.20	0.13	0.17	0.18	0.15	0.16	0.20	0.13

Fuente: elaboración propia.

3.2 Resultados para editoriales

La Figura 4 presenta los porcentajes de reducción de palabras obtenidos con los textos de editoriales. Similar a los resultados obtenidos para las noticias, los resúmenes, empleando el método basado en la matriz de similitud, presentan la reducción más alta de palabras. Para los métodos TF-IDF y TF, en la mayoría de los documentos la reducción de palabras está por debajo de la obtenida por el resumen de referencia. Los valores promedio de reducción son comparables a los obtenidos para las noticias: 70 % de reducción de palabras para el método basado en similitud, 52 % para TF-IDF, 42 % para TF y 60 % para el resumen de referencia.

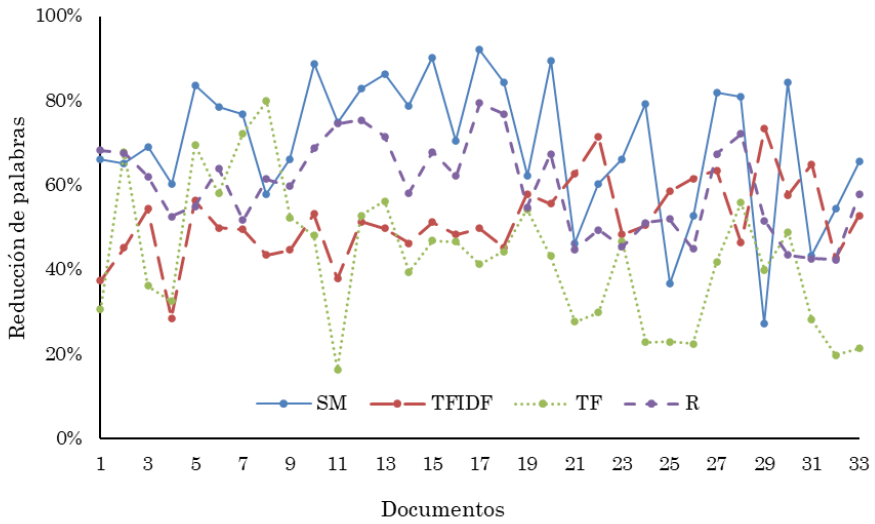

FIGURA 4.
Porcentaje de reducción de palabras en resúmenes de las editoriales empleando el método basado en matriz de similitud (SM), las métricas TF-IDF y TF, y el resumen de referencia (R)
Fuente: elaboración propia.

La Tabla 3 presenta los resultados obtenidos para los textos de editoriales empleando la métrica ROUGE-1. De acuerdo con la precisión, el mejor resumen fue obtenido por el método basado en la matriz de similitud con un 69 %, similar a lo obtenido para las noticias. En términos de sensibilidad, el mejor resultado lo presenta el algoritmo basado en TF con un 77 %. El valor F más alto es obtenido con el algoritmo TF (63 %). Tanto para la precisión como para el valor F, las desviaciones estándar se encuentran por debajo de 15 %, pero en el caso de la sensibilidad llegan a ser tan altas como 19 %. En estos resultados se notó cómo la matriz de similitud cuenta con los mejores puntajes para R con un 19 % y para F con un 14 %. Los resultados obtenidos a partir de la métrica ROUGE-L se presentan en la Tabla 4. En este caso, de acuerdo con la precisión, tanto el método TF, como el TF-IDF, obtienen 45 % de precisión. El método TF obtiene el mejor resultado de

acuerdo con la sensibilidad (63 %) y valor F (51 %). Las desviaciones estándar varían entre 9 % y 18 % para todas las métricas.

TABLA 3.
Promedio (μ) y desviación estándar (σ) para las métricas ROUGE-1 para los resúmenes automáticos de noticias empleando los algoritmos basados en similitud (SM), TF-IDF y TF.

	SM			TF-IDF			TF		
	P	R	F	P	R	F	P	R	F
μ	0.69	0.49	0.55	0.58	0.70	0.61	0.56	0.77	0.63
σ	0.09	0.19	0.14	0.13	0.15	0.08	0.12	0.14	0.09

Fuente: elaboración propia.

TABLA 4.
Promedio (μ) y desviación estándar (σ) para las métricas ROUGE-L para los resúmenes automáticos de noticias empleando los algoritmos basados en similitud (SM), TF-IDF y TF.

	SM			TF-IDF			TF		
	P	R	F	P	R	F	P	R	F
μ	0.41	0.30	0.33	0.45	0.53	0.47	0.45	0.63	0.51
σ	0.09	0.15	0.11	0.14	0.13	0.10	0.12	0.18	0.12

Fuente: elaboración propia.

3.3 Análisis y discusión de resultados

Los resultados evidencian que el método basado en la matriz de similitud y el TF son los mejores en términos generales, pero la matriz de similitud obtiene el mejor desempeño de acuerdo con el porcentaje de reducción de palabras y el promedio de precisión de las métricas ROUGE-1 y ROUGE-L. Este método identifica las frases más relevantes a partir de una métrica de similitud basada en la distancia coseno. Las N frases con los pesos más altos son empleadas para construir el resumen, en este caso N se fijó a cinco. Esta es la razón principal por la cual el método basado en la matriz de similitud obtiene el mejor desempeño en porcentaje de reducción de palabras. Por su parte, la precisión es una medida de la relevancia de las palabras (ROUGE-1) o frases (ROUGE-L) incluidas en el resumen, relevancia que se determina por el porcentaje de palabras en común entre el resumen automático y el de referencia, y el número de palabras en el resumen automático. Los resultados obtenidos con esta técnica (SM), tanto para noticias y editoriales, son similares. Sin embargo, la variabilidad de la precisión para las noticias (16 % para ROUGE-1 y 13 % para ROUGE-L) es más alta que para los textos editoriales (9 % para ROUGE-1 y 9 % para ROUGE-L, respectivamente).

A diferencia de los textos argumentativo de las editoriales, las noticias presentan estilos de escritura más variado y menos formal. Por ejemplo, el siguiente texto corresponde al resumen automático obtenido por el método basado en la matriz de similitud para una noticia con un porcentaje de precisión del 24.8 %:

> *Esto, teniendo en cuenta que, si bien el Ministerio de Salud trabaja en un protocolo específico para moldear la apertura de bares de una manera segura, aún no se han expedido. Sin alcohol, otro de los puntos que se destacan con respecto a la posibilidad de hacer estas pruebas es que, por lo pronto, no se podrá consumir bebidas alcohólicas dentro de estos negocios, según aclaró el Ministerio. Pese a esto, Asobares manifestó que se encuentran trabajando para que se avale de manera gradual, desde septiembre. La idea de volver a escuchar música y hablar con amigos en un bar cada vez está más cerca, pues, según confirmó el Ministro de Salud, Fernando Ruiz, los gobernantes locales ya están en capacidad de solicitar al Ministerio del Interior la apertura de estos negocios, a manera de pruebas piloto, luego de que dicha cartera emitiera una circular al respecto, en la que se da esta posibilidad no solo a municipios sin covid-19 o con baja afectación, sino también a aquellos con una mayor presencia del virus. Juan Pablo Valenzuela, presidente del gremio de Antioquia, manifestó que por ahora "en el empresariado no genera mucha expectativa, teniendo en cuenta que no es sostenible la apertura".*

El resumen de referencia para esta noticia se presenta a continuación:

> *Según el Ministro de Salud, Fernando Ruiz, los gobernantes locales pueden solicitar al Ministerio del Interior la apertura de bares. Los negocios que se abran deberán cumplir con protocolos de bioseguridad. No se podrá consumir bebidas alcohólicas dentro de estos negocios. Asobares manifestó que se encuentra trabajando para que se avale de manera gradual, desde septiembre.*

En este caso se puede observar como el algoritmo basado en la matriz de similitud no sintetiza completamente cada frase. Por ejemplo, la última frase del párrafo del resumen automático incluye 117 palabras.

Por otra parte, de acuerdo con la medida de sensibilidad, el mejor resultado lo obtuvo el algoritmo basado en la frecuencia de las palabras TF, para ambos tipos de documento. En el caso de noticias obtuvo 69 % para ROUGE-1 y 54 % para ROUGE-L, para las editoriales estos porcentajes fueron 77 % y 63 %, respectivamente. La sensibilidad de las métricas ROUGE-1 y ROUGE-L se relacionan con la retención de la información que se encuentra en el resumen de referencia. A saber, para la noticia del ejemplo anterior, el algoritmo TF obtiene el siguiente resumen:

> *La idea de volver a escuchar música y hablar con amigos en un bar cada vez está más cerca, pues, según confirmó el Ministro de Salud, Fernando Ruiz, los gobernantes locales ya están en capacidad de solicitar al Ministerio del Interior la apertura de estos negocios, a manera de pruebas piloto, luego de que dicha cartera emitiera una circular al respecto, en la que se da esta posibilidad no solo a municipios sin covid-19 o con baja afectación, sino también a aquellos con una mayor presencia del virus. Según Carlos Agudelo, epidemiólogo e infectólogo de la Clínica Universitaria Bolivariana, "desde el punto de vista económico el tema de los bares y restaurantes es crítico, pero preocupa que los casos aumenten, porque mantener el distanciamiento social y el uso de la mascarilla es difícil en estos establecimientos, y más ahora que hay ciudades como Medellín y Bogotá que están en el pico".*

Estos resultados demuestran que, aunque en caso de precisión, sensibilidad y reducción de palabras se puede hacer una comparación cuantitativa del desempeño de los algoritmos existentes, aún hace falta mejorar la obtención de resúmenes automáticos. Una forma de hacerlo es a través de la inclusión de la semántica y el contexto de la aplicación. Sin embargo, a diferencia de los algoritmos no supervisados empleados en este artículo, los modelos de resumen automáticos que se basan en la semántica o en conceptos son usualmente supervisados, requiriendo un entrenamiento previo.

4. CONCLUSIONES

Este articulo presentó la comparación de tres métodos extractivos de resumen automático de textos aplicados al procesamiento de editoriales y noticias. El primer método se basa en una matriz de similitud para calcular un peso de cada frase en el documento. Por su parte, los otros métodos se basan en el cálculo de la frecuencia de las palabras para determinar su importancia. La comparación de los algoritmos se realizó empleando el porcentaje de reducción de palabras y la métrica ROUGE, calculada a partir de los resúmenes automáticos y un resumen de referencia.

Los resultados permiten identificar que, en términos de porcentaje de reducción de palabras y precisión (relevancia de las palabras o frases incluidas en el resumen), el método que permite obtener los mejores resultados, tanto para noticias como para editoriales, es el basado en la matriz de similitud. Sin embargo, en términos de sensibilidad, el cual determina qué tan cercano es el resumen al texto de referencia, los resultados evidencian un mejor desempeño para el método basado en la frecuencia de las palabras TF.

Es importante señalar la limitación de este estudio, en cuanto los resultados dependen del resumen de referencia, que a su vez es una construcción sujeta a la interpretación del que lo construye. Para reducir esta limitación, se determinó que el resumen de referencia incluyera cierta información de acuerdo con el tipo de documento: para las editoriales, esta información correspondía a la problemática y opinión del autor; para las noticias, el suceso principal, donde y cuando ocurrió, y quiénes estaban involucrados. Adicionalmente, este artículo no incluye una comparación del efecto que tienen los parámetros de los algoritmos, específicamente

no se estudia cómo afecta la selección de N (número de frases con mayor peso) en el desempeño de los algoritmos.

Se destaca que los métodos empleados son extractivos y no supervisados. Por tanto, dependen de métricas que miden la importancia de las frases o palabras en el texto (similitud o frecuencia en las técnicas seleccionadas) sin un entrenamiento previo. Por lo anterior, son métodos que no tienen en cuenta el contexto o tipo de documento para el procesamiento de los datos. En consecuencia, como trabajo futuro se propone explorar técnicas que incluyan el contexto y la aplicación para mejorar los resultados.

Agradecimientos

El desarrollo de este trabajo de investigación fue financiado por la Convocatoria Jóvenes Investigadores e Innovadores para Grupos de Investigación ITM-2020, del Instituto Tecnológico Metropolitano.

REFERENCIAS

[1] K. R. Chowdhary, "Natural language processing," en *Fundamentals of Artificial Intelligence*, New Delhi: Springer, 2020, pp- 603-649. https://doi.org/10.1007/978-81-322-3972-7_19

[2] A. Cortez Vásquez, H. Vega Huerta, J. Pariona Quispe, A. M. Huayna, "Procesamiento de lenguaje natural", *Revista de Investigación de Sistemas e Informática*, vol. 6, no. 2, pp. 45-54, dic. 2009. https://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/5923

[3] A. Gelbukh, "Procesamiento de Lenguaje Natural y sus Aplicaciones", *Komputer Sapiens*, vol. 1, pp. 6-11, jun. 2010. https://www.gelbukh.com/CV/Publications/2010/Procesamiento%20de%20lenguaje%20natural%20y%20sus%20aplicaciones.pdf

[4] A. Rivera Arrizabalaga, S. Rivera Velasco, "Origen del lenguaje: un enfoque multidisciplinar", *Ludus Vitalis*, vol. 17, no. 31, pp. 103-141, 2009. http://ludus-vitalis.org/ojs/index.php/ludus/article/view/277

[5] V. Gupta, G. S. Lehal, "A Survey of Text Mining Techniques and Applications", *Journal of Emerging Technologies in Web Intelligence*, vol. 1, no. 1, pp. 60-76, Aug. 2009. http://learnpunjabi.org/pdf/gslehal-pap18.pdf

[6] S. Naqeeb Khan, N. Mohd Nawi, M. Imrona, A. Shahzad, A. Ullah, A. Ur- Rahman, "Opinion Mining Summarization and Automation Process: A Survey", *International Journal on Advanced Science Engineering Information Technology*, vol. 8, no. 5, pp. 1836-1844, 2018. http://dx.doi.org/10.18517/ijaseit.8.5.5002

[7] C. Yew-Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", In *Text summarization branches out, Association for Computational Linguistics*, pp. 74-81, 2004. https://www.aclweb.org/anthology/W04-1013.pdf

[8] Z. Li, Z. Peng, S. Tang, C. Zhang, H. Ma, "Text Summarization Method Based on Double Attention Pointer Network", *IEEE Access*, vol. 8, pp. 11279-11288, Jan. 2020. https://doi.org/10.1109/ACCESS.2020.2965575

[9] M. González Boluda, "Estudio comparativo de traductores automáticos en línea: Systran, reverso y google", *Núcleo*, vol. 22, no. 27, pp. 187-216, dic. 2010. http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S0798-97842010000100008

[10] A. Hernández Castañeda, R. A. García Hernández, Y. Ledeneva, C. E. Millán Hernández, "Extractive Automatic Text Summarization Based on Lexical-Semantic Keywords", *IEEE Access*, vol. 8, pp. 49896-49907, Mar. 2020. https://doi.org/10.1109/ACCESS.2020.2980226

[11] S. Kumar Saha, D. Rao Ch., "Development of a practical system for computerized evaluation of descriptive answers of middle school level students." *Interactive Learning Environments*, pp. 1-14, Ago. 2019. https://doi.org/10.1080/10494820.2019.1651743

[12] J. Rose, C. Lennerholt, "Low-cost text mining as a strategy for qualitative researchers", *Electronic Journal of Business Research Methods*, vol. 15, no. 1, pp. 2-16, Apr. 2017. https://www.researchgate.net/publication/315702194_Low_cost_text_mining_as_a_strategy_for_qualitative_researchers

[13] G. A. Matias Mendoza, Y. Ledeneva, R. A García Hernández, "*Detección de ideas principales y composición de resúmenes en inglés, español, portugués y ruso. 60 años de investigación*", Alfaomega Grupo Editor, S.A. 2020. https://www.semanticscholar.org/paper/Detecci%C3%B3n-de-ideas-principales-y-composici%C3%B3n-de-en-Mendoza-Ledeneva/4ae110ed12c30b76a869206092b097605ffc4f56

[14] M. D. Bustamante-Rodríguez, A. A. Piedrahita-Ospina, I. M. Ramírez-Velásquez, "Modelo para detección automática de errores léxico-sintácticos en textos escritos en español", *TecnoLógicas*, vol. 21, no. 42, pp. 199-209, May. 2018. https://doi.org/10.22430/22565337.788

[15] R. Elbarougy, G. Behery, A. El Khatib, "Extractive Arabic Text Summarization Using Modified PageRank Algorithm", *Egyptian Informatics Journal*, vol. 21, no. 2, pp. 73-81, Jul. 2020. https://doi.org/10.1016/j.eij.2019.11.001

[16] R. Chandra Belwal, S. Rai, A. Gupta. "A new graph-based extractive text summarization using keywords or topic modeling." *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-16, Oct. 2020. https://doi.org/10.1007/s12652-020-02591-x

[17] J. Steinberger, K. Ježek, "Evaluation measures for text summarization", *Computing and Informatics*, vol. 28, no. 2, pp. 251–275. Mar. 2009. https://cai.type.sk/content/2009/2/evaluation-measures-for-text-summarization/1726.pdf

[18] H. Christian, M. Pramodana Agus, D. Suhartono, "Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)", *ComTech: Computer, Mathematics and Engineering Applications*, vol. 7, no. 4, pp. 285-294, Dic. 2016. https://doi.org/10.21512/comtech.v7i4.3746

[19] I. Manterola, A. Diaz de Ilarraza, K. Gojenola, K. Sarasola, "Recursos en euskera para la herramienta NLTK para enseñanza de procesamiento del lenguaje natural." *Procesamiento del Lenguaje Natural*, no. 45, pp. 305-306, Sep. 2010. https://www.redalyc.org/pdf/5157/515751745045.pdf

Notas

- **CONFLICTOS DE INTERÉS**
  Los autores declaran no tener ningún conflicto de interés.
- **CONTRIBUCIÓN DE LOS AUTORES**
  S. López-Trujillo participó del diseño metodológico de la investigación, construcción de la base de datos, implementación de los algoritmos, obtención de resultados, análisis y escritura del manuscrito. M.C. Torres-Madroñero participó del diseño metodológico de la investigación, construcción de la base de datos, análisis y discusión de resultados, revisión y escritura del manuscrito.

Información adicional

Cómo citar / How to cite: S. López-Trujillo; M. C. Torres-Madroñero, "Comparación de algoritmos de resumen de texto para el procesamiento de editoriales y noticias en español", *TecnoLógicas*, vol. 24, nro. 51, e1816, 2021. https://doi.org/10.22430/22565337.1816