ISSN 1870-4069

Selección de características de representaciones de texto de BETO usando un algoritmo genético

Juan José Guzmán-Landa, José Clemente Hernández-Hernández,
Guillermo de Jesús Hoyos-Rivera, Efrén Mezura-Montes

Instituto de Investigaciones en Inteligencia Artificial,
Universidad Veracruzana,
México

zS21000453@estudiantes.uv.mx,
jclementehdzhdz@gmail.com, {ghoyos, emezura}@uv.mx

Resumen. El Procesamiento del Lenguaje Natural es un área que se está volviendo sumamente importante dentro de la investigación en Inteligencia Artificial, esto incluye el análisis de sentimientos, la traducción automática, y la generación de texto, entre otros. El análisis manual sobre el texto es un desafío significativo debido a la gran cantidad de datos que se generan a través de los medios sociales en la red; el análisis asistido por computadora es una opción viable. Recientemente, han emergido múltiples tareas para el manejo automático del texto, especialmente en el idioma inglés de las tareas previamente mencionadas. Los casos más representativos se encuentran dentro de las técnicas de Aprendizaje Profundo, específicamente aquellos relacionados con el modelo BERT y otros modelos de tipo Transformer. Dicho modelo genera un vector de 768 características para representar cada palabra o fragmento de palabra de manera numérica. El número de características usualmente tiene una ausencia de justificación y de descripción. Además, la mayoría de los trabajos de investigación en la clasificación de sentimientos se encuentran fuera del idioma español, y del uso de técnicas con modelos de tipo Transformer. Basado en lo antes mencionado, este trabajo propone hacer uso de un conjunto de datos en español, a través de un Algoritmo Genético para la selección de características en un enfoque de envoltura, con el fin de reducir el número de características de los vectores generados por un modelo entrenado en conjuntos de datos en español, BETO (BERT para el español), y obtener un subconjunto de ellas, asimismo, verificar si con el nuevo conjunto de características se puede obtener un buen desempeño en la clasificación de sentimientos. Los resultados obtenidos en una serie de experimentos indican un desempeño competitivo, en la tarea de clasificación de sentimientos, incluso con una representación mucho menor con respecto al total de las características originales.

Palabras clave: BERT, algoritmo genético, representaciones vectoriales, selección de características, reducción de dimensionalidad.

Feature Selection of BETO Token Embeddings Using a Genetic Algorithm

Abstract. Natural Language Processing is an area that is becoming increasingly important in Artificial Intelligence research, including sentiment analysis,

machine translation, and language understanding, among others. Handcrafted analysis over text is very challenging due to the large amounts of data generated through digital social networks; computer-assisted analysis is a viable option. Recently, multiple text-handling task proposals have emerged to tackle various assignments automatically, especially in the English language of the aforementioned tasks. Most representative cases are found in the Deep Learning techniques, specifically those related to BERT and other Transformer models. Such a model generates a continuous vector of 768 elements representing a word or a token. The number of characteristics usually has an absence of justification and description. In addition, most of the research works on sentiment classification are outside of the Spanish language and the use of Transformer-type modeling techniques. Based on the above, this work proposes to make use of a Spanish data set through a Genetic Algorithm for feature selection in a wrapper approach in order to reduce the number of features of the embeddings generated by BETO (BERT for Spanish), and to obtain a subset of them, also, to verify if with the new set of features, a good performance in sentiment classification can be obtained. The results obtained in a series of experiments indicate a competitive performance, in the sentiment classification task, even with a much smaller representation with respect to the total of the original features.

Keywords: BERT, genetic algorithm, token embeddings, feature selection, dimensionality reduction.

1. Introducción

El texto, visto por una máquina, es una secuencia de símbolos que no tienen significado alguno, dado que interpretar este tipo de recursos es una actividad inherentemente humana.

El texto puede ser fácilmente manipulado en diferentes formas con una máquina, por ejemplo, dividir el texto por caracteres, detectar espacios en blanco y saltos de línea, entre otros; también se puede cambiar la posición de los caracteres, reemplazar y cambiar el uso de mayúsculas y minúsculas, pero nada que pueda dar un significado al texto. Todo lo anterior es meramente sintáctico, y centrado en alteraciones morfológicas.

BERT [1] es una opción que permite ir más a profundidad de las tareas mencionadas previamente. Usando este modelo, se obtiene una representación vectorial del texto, es decir, el texto cambia a una forma numérica, por lo tanto, se cuenta con un procesamiento computacional más adecuado con respecto al significado del texto.

BERT también tiene una variante para el español llamado BETO [4]. Este último obtiene una representación numérica del texto, sólo que con un contexto y significado diferente debido al idioma usado.

La representación resultante es habitualmente generada usando técnicas de Aprendizaje Profundo (AP). En este caso en especial, el modelo está basado en el mecanismo de atención [11], el cual genera buenos resultados en un amplio rango de tareas.

Aún así, tiende a requerir bastantes recursos computacionales para el procesamiento. Por ejemplo, una oración de tan solo 50 palabras genera una matriz de 50×768, donde 768 es el número de valores continuos que representan a una palabra o token (fragmento de palabra) [13], y por lo tanto, esto se convierte en una entrada de 38,400 características. Este resultado puede impactar en el número de cálculos cuando el conjunto de oraciones es mayor.

Por esta razón, reducir el número de características que representan al texto se convierte en una tarea importante, eventualmente esperando lograr un similar, o incluso mejor desempeño que la representación original, al momento de realizar tareas de aprendizaje como la clasificación.

Este trabajo de investigación propone un enfoque de envoltura usando un Algoritmo Genético (AG) como un algoritmo de búsqueda y una Red Neuronal Artificial (RNA) como un clasificador para realizar la tarea de Selección de Características (SC), todo esto utilizando el conjunto de datos de reseñas de películas IMDb[1] para el idioma en español. La tarea de evaluación se centra en realizar Análisis de Sentimientos (AS) de dicho conjunto de datos.

Este artículo tiene como contribución, además de mostrar un estudio experimental sobre la reducción de características de las representaciones generadas por BETO, demostrar, que sólo algunas características son necesarias para resolver una tarea en específico, que para el caso de estudio actual es el AS.

El resto del trabajo de investigación está organizado de la siguiente forma: en la Sección 2, se presentan los trabajos relacionados con este estudio; en la Sección 3, se presenta en detalle el enfoque propuesto. La Sección 4 incluye la descripción de los datos utilizados, mientras que la Sección 5 muestra los experimentos y sus correspondientes resultados. Finalmente, en la Sección 6, se presentan las conclusiones y el trabajo futuro.

2. Trabajos relacionados

Como parte de los antecedentes, esta Sección describe algunos de los principales trabajos de investigación relacionados en el proceso de SC sobre representaciones vectoriales del texto.

En [10], se describe un proceso de SC con una representación de texto de tipo, Frecuencia de Termino - Frecuencia Inversa de Documento, por su siglas en inglés, TF-IDF. El mecanismo de SC reduce la dimensionalidad de la representación a través de un AG y el método de Análisis de Componentes Principales, por sus siglas en inglés, PCA.

Este último obtiene una representación a nivel de documentos para una tarea en específico, lo cual no permite utilizar esa representación a nivel de palabras, y por lo tanto, no puede ser orientado para otras tareas. Para trabajar otro tipo de tarea dentro del PLN, se rquiere otro proceso de SC. Los resultados finales muestran que después de correr el mecanismo de SC, se obtiene una mejor precisión de clasificación.

En [3], se propone mejorar la precisión de clasificación de textos de diagnósticos médicos mediante una SC con un AG para la reducción de dimensionalidad.

[1] https://www.kaggle.com/datasets/luisdiegofv97/imdb-dataset-of-50k-movie-reviews-spanish

Algorithm 1: Algoritmo Genético basado en envoltura para la Selección de Características de BETO

Data: Reseñas de películas representada por BETO
Result: El subconjunto de características con la mejor precisión de clasificación
$P \leftarrow$ Inicializar una población;
Calcular la aptitud de cada solución en P;
while *MAX_GEN no es alcanzado* **do**
    Seleccionar T soluciones de P usando la Selección por Torneo;
    Aplicar cruza a la solución en T;
    Mutar al descendiente generado después de la cruza;
    Calcular la aptitud de cada descendiente;
    Aplicar reemplazo de soluciones;
end

En este trabajo de investigación, la representación del texto no se especifica con claridad; se puede intuir que se utilizó una bolsa de palabras, conocido por sus siglas en inglés, BoW. Los resultados mostrados concluyen en una buena reducción de la dimensionalidad.

El trabajo de investigación descrito en [9] propone un mecanismo de SC para reducir la dimensionalidad de los vectores generados por el modelo Word2Vec. El proceso principal consiste en filtrar las características que destacan más de una categoría y están ausentes en otras. El método producido se compara con mecanismos de SC tradicionales. Lo resultados finales muestran un desempeño similar en ambos casos.

Los autores de [12] propusieron un mecanismo de SC antes de representar el texto a través de un modelo, que en este caso es BERT. Por ejemplo, obteniendo las palabras más representativas de un texto a través de un método como TF-IDF. Dado que BERT no se desempeña muy bien cuando la longitud del texto es muy grande, el proceso de SC descrito es benéfico para la clasificación de textos grandes.

A partir de la revisión de la literatura anterior, se puede observar que los trabajos de investigación se enfocan particularmente en BERT y Word2Vec, dejando sin explorar, de acuerdo a la revisón hecha por los autores, la rama del español que usa BETO y todo lo que eso implica.

Motivado por lo antes mencionado, este trabajo de investigación introduce un método de SC para las representaciones vectoriales de BETO, un modelo entrenado para el idioma español.

3. Enfoque propuesto

Para lograr el propósito de este trabajo de investigación, se consideró una SC basada en envoltura con un AG [14]. En este sentido el proceso de SC de tipo envoltura, incluye un clasificador que servirá como función de aptitud para evaluar la calidad de las características seleccionadas, mediante la precisión del modelo. Por otra parte, el proceso de un AG está basado en la evolución natural de las especies, y su proceso se puede observar en el Algoritmo 1.

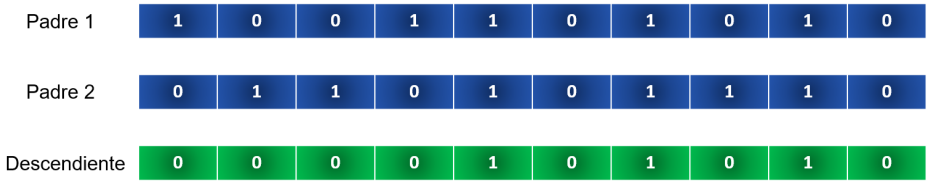

Fig. 1. Operador de cruza AND.

Fig. 2. Mutación simple modificada.

La complejidad del problema a resolver tiende a aumentar cuando el número de características disponible es grande. En el caso de estudio actual, se tienen 768 características, un número alto si una solución algorítmica de fuerza bruta trata de reducir esa dimensionalidad [6].

Por esto mismo, un AG es viable. Además, de que es posible que exista más de una solución buena para el problema en cuestión. Un AG explora el espacio de búsqueda y explota las zonas donde se encuentran mejores soluciones y, así, lograr una mejor calidad en la selección o reducción de características.

Una población inicial P de soluciones potenciales son generadas de manera aleatoria y evolucionadas durante un número específico de generaciones, usando operadores de variación como la cruza y mutación, los cuales están completamente ligados a la representación de la solución.

Después de algunas generaciones, es posible encontrar una solución competitiva con respecto a la función de aptitud, lo que representa la calidad de la solución del problema a resolver.

El AG utiliza una representación con cadenas binarias de tamaño igual al total de características de las representaciones vectoriales del conjunto de datos de interés, para definir una solución potencial.

Dicha representación indica qué características son seleccionadas o descartadas, por ejemplo, “1” indica que una característica será seleccionada, mientras que un “0” significa lo opuesto. Las características seleccionadas en esta cadena binaria vienen a partir de las representaciones vectoriales continuas generadas por BETO para cada palabra o token visto en el corpus de texto, que para el caso de BETO es 768, para el caso de Word2Vec, normalmente, es 300.

Considerando el enfoque de envoltura, un clasificador se utiliza como parte de la función de aptitud en el AG. Una solución potencial X es evaluada con la Ecuación 1:

$$aptitud(X) = w_1 * C(X) + w_2 * \frac{|X| - R(X)}{|X|}, \tag{1}$$

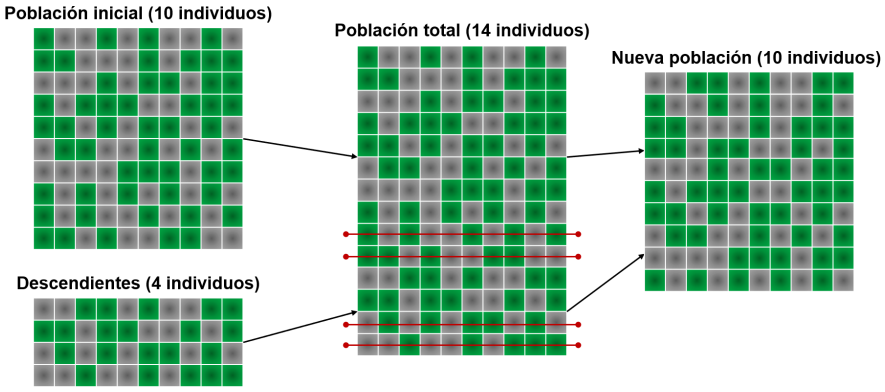

Fig. 3. Reemplazo basado en la aptitud; donde se puede ver las características seleccionadas en color verde y cada individuo representado por las filas de las matrices.

donde w_1 y w_2 son los pesos de importancia de cada factor en la función de aptitud, siendo $w_1 + w_2 = 1$. $C(X)$ es la precisión de clasificación usando las características seleccionadas en X, mientras que $R(X)$ es el número actual de las características seleccionadas en la solución X.

Es importante enfatizar que la polaridad de la clasificación es considerada a partir de las reseñas de películas en español, usando la RNA generada en el trabajo de investigación descrito en [5]. La descripción de dicha RNA se encuentra en la Sección 5.

La técnica de selección de padres adoptada en este trabajo es la selección mediante torneo. En este caso, el tamaño del torneo es definido por T, que es el número de soluciones tomadas de manera aleatoria de la población, de donde se selecciona la mejor solución, y además se considera realizar, posteriormente, la operación de cruza. El proceso de selección por torneo es llamada n veces, donde n es un número par.

Con base en la probabilidad de cruza, un par de padres son recombinados por un operador de cruza inspirado en el operador lógico AND, el cual demostró resultados competitivos en problemas de SC [3]. La Figura 1 detalla este operador.

Después de la cruza, se aplica una versión modificada de la mutación simple, con una probabilidad de mutación para cada descendiente que es definida previamente. Inspirado por el operador de mutación simple para la codificación de cadenas binarias, donde una simple posición está sujeta a la operación de bitflip, donde el valor de “1” se cambia por un “0” y vice-versa, el operador de mutación modificado usado aquí, siempre selecciona aleatoriamente una posición con un “1” y la cambia por un “0”, así como también, una posición con un “0” es aleatoriamente seleccionada y cambiada por un “1”.

De esta forma, el operador de mutación cambia las características seleccionadas, pero mantiene el número de características seleccionadas. En contraste, el operador de cruza es capaz de reducir el número de características. Un ejemplo visual del operador de mutación modificado puede ser visto en la Figura 2.

El último paso del AG se encuentra en el reemplazo (también conocido como selección de supervivencia o selección ambiental) para mantener el tamaño de la población fijo después de la creación de los descendientes.

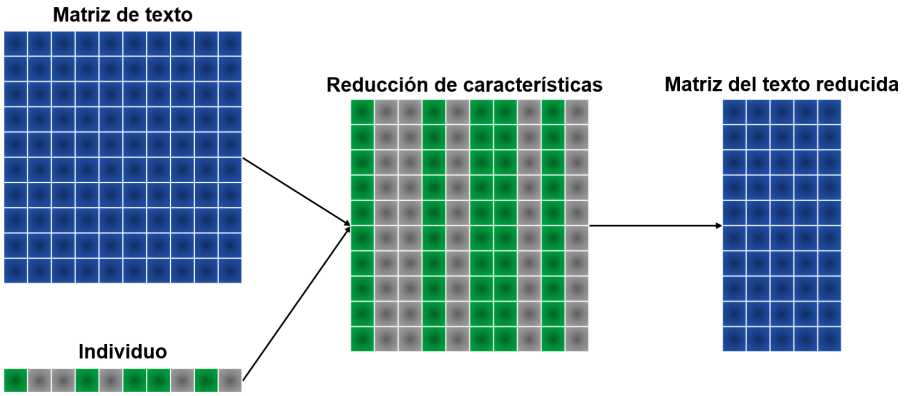

Fig. 4. Selección de Características con un individuo del AG.

A diferencia del reemplazo generacional tradicional en un AG canónico, donde la población actual es descartada, y todos los descendientes conforman la población de la siguiente generación, el proceso está inspirado por el reemplazo $(\mu + \lambda)$ de las Estrategias Evolutivas, donde la población actual y los descendientes se unen en un sólo conjunto, y los mejores $|P|$ individuos, basados en el valor de aptitud, se mantienen en la población para la siguiente generación, siendo $|P|$ el tamaño de la población original. En la Figura 3, se detalla este reemplazo.

4. Pre-procesamiento de los datos

Los datos utilizados en este estudio provienen del repositorio Kaggle, el cual contiene un conjunto de reseñas de películas en español IMDb. El corpus original está escrito en Inglés [7], y fue traducido por [2].

Dicha base de datos consiste de 50,000 reseñas de películas y es usada para la clasificación de sentimientos en dos clases; tiene 25,000 reseñas negativas y 25,000 positivas.

Un breve estudio del número de palabras en el corpus se llevó a cabo. Con ello se pudo concluir que existe una alta concentración de reseñas que tienen entre 125 y 375 palabras. Por esto mismo, el subconjunto de reseñas extraídas contienen entre 200 y 400 palabras, un porcentaje de ellas fueron utilizadas para evaluar el rendimiento del AG descrito en la sección anterior. El resultado de la extracción fue de 7,000 reseñas por cada clase, negativa y positiva.

Para tener los datos listos, fue necesario someter cada reseña a un proceso que consiste en pasar todo el texto a minúsculas y representar cada una de las palabras o tokens de las reseñas, en su respectiva representación vectorial generada por BETO. Lo anterior se realizó usando la biblioteca Bert-Tokenizer, que implementa el algoritmo de tokenización WordPiece [13], proporcionada por el lenguaje de programación Python.

Con los datos listos, el AG descrito en las secciones previas puede ser probado. En cuanto a la evaluación realizada por la función de aptitud en el enfoque basado en envoltura, el siguiente proceso se lleva a cabo para cada individuo de la población:

Tabla 1. Parámetros usados en cada experimento.

Parámetro	**Exp. 1**	**Exp. 2**	**Exp. 3**
Tamaño de la población	50	50	-
Número de reseñas (r)	40	100	100
MAX_GEN	10	10	-
Épocas de entrenamiento (e)	5	5	30
Palabras por reseña	30	50	50
w_1	0.2	0.2	-
w_2	0.8	0.8	-
Tamaño del torneo (T)	2	2	-
Número de padres (n)	20	20	-
Ejecuciones del AG	10	10	30

1. Las representaciones vectoriales generadas por BETO se van a concentrar en una matriz, siendo sus columnas, las que se van a seleccionar tomando en cuenta los valores de un individuo X, tal y como se explicó en la sección anterior.
2. Una RNA se utiliza como clasificador. Una solución X de la población del AG, con su correspondiente conjunto de características, entrena al clasificador durante un número de épocas e para un número de reseñas r, representado con BETO. Un ejemplo de la selección de características a partir de un individuo del AG se puede visualizar en la Figura 4, con una sola reseña. El proceso se realiza con todas las reseñas obtenidas después del estudio del número de palabras.
3. La RNA se entrena usando un subconjunto de las reseñas, tal y como fue descrito en la Sección 3. La red es evaluada usando otro subconjunto de las reseñas, y con ello se obtiene el rendimiento de la precisión.

5. Experimentos y resultados

Tres experimentos se llevaron a cabo para evaluar el desempeño del AG propuesto: (1) calibrar los parámetros para los operadores de cruza y mutación, (2) analizar la habilidad de la propuesta para reducir el número de características seleccionadas, y (3) comparar el desempeño de la RNA usando el número de características reducido contra el número de características original.

Todos los experimentos fueron realizados usando Python en su versión 3.10 y una computadora con Intel Xeon(R) CPU E542 de 8 núcleos, 2.80GHz, 6GB RAM, y Ubuntu 22.04.1 LTS. Es importante resaltar que todos los experimentos fueron limitados debido al recurso computacional disponible. Por lo tanto, el AG funciona con una porción de los datos seleccionados (ver Sección 4).

El número de reseñas y palabras usado en el experimento 1 es menor debido a su alto costo computacional. En la Tabla 1 se muestran los parámetros utilizados en cada experimento. La RNA adoptada en la función de evaluación del AG tiene la siguiente configuración: (1) una capa de entrada, con un número de neuronas a partir del producto entre la longitud del número de tokens usados y el número de características, e. g., 50 tokens $\times$ 768 o cualquier otro número características encontradas, y (2) dos neuronas como capa de salida, análogo con el número de clases del conjunto de datos.

Tabla 2. Experimento 1. Resultado de 10 ejecuciones del AG para cada combinación de valores de parámetros y así obtener las mejores probabilidades de cruza y mutación (%).

% de cruza	**% de mutación**	**Características (Avg, Max, Min, Med)**				**Precisión (Avg, Max, Min, Med)**				**Aptitud (Avg, Max, Min, Med)**			
0.2	0.2	5.1	6	1	6	0.78	0.83	0.75	0.75	0.951	0.965	0.943	0.945
0.2	0.4	4.3	6	1	6	0.80	0.91	0.75	0.79	0.955	0.982	0.943	0.954
0.2	0.6	8	37	1	6	0.82	0.91	0.75	0.83	0.956	0.977	0.943	0.960
0.2	0.8	8.8	37	3	6	0.79	0.91	0.75	0.75	0.949	0.960	0.943	0.944
0.4	0.2	4.4	6	1	5	0.80	0.91	0.75	0.79	0.955	0.979	0.943	0.954
0.4	**0.4**	5.8	15	1	6	0.82	0.91	0.75	0.83	**0.958**	0.977	0.945	0.960
0.4	0.6	4.5	14	1	4	0.75	0.83	0.66	0.75	0.945	0.963	0.929	0.944
0.4	0.8	7.5	15	5	6	0.80	0.91	0.75	0.79	0.952	0.967	0.943	0.948
0.6	0.2	3.9	6	1	4	0.76	0.83	0.75	0.75	0.949	0.960	0.943	0.946
0.6	0.4	4.5	6	1	6	0.80	0.83	0.75	0.83	0.955	0.964	0.943	0.960
0.6	0.6	4.8	15	1	4	0.77	0.83	0.75	0.75	0.950	0.965	0.934	0.946
0.6	0.8	6.2	15	1	5	0.79	0.83	0.75	0.79	0.951	0.963	0.943	0.950
0.8	0.2	10.3	19	1	7.5	0.84	1	0.75	0.83	0.957	0.981	0.943	0.958
0.8	0.4	7.1	19	1	3.5	0.79	1	0.75	0.75	0.950	0.980	0.934	0.948
0.8	0.6	4.5	15	1	3.5	0.74	0.83	0.66	0.75	0.943	0.960	0.929	0.945
0.8	0.8	3	10	1	2.5	0.73	0.75	0.66	0.75	0.943	0.948	0.931	0.946

Para el entrenamiento de la RNA se usa el optimizador Adam, la capa de salida utiliza como función de activación la función softmax, y la función de pérdida es, por su nombre en inglés, Negative Log Likelihood.

Considerando el primer experimento, se llevó a cabo una calibración de parámetros para obtener la configuración más adecuada para los operadores de cruza y mutación. Ambas probabilidades de cruza y mutación fueron variadas entre 0.2 y 0.8, y se ejecutaron 10 veces para cada combinación.

Para generar los resultados mostrados en la Tabla 2, se realizó lo siguiente: (1) para cada ejecución el mejor individuo fue seleccionado, y su correspondiente aptitud, número de características, y la precisión del clasificador fue recuperada; (2) usando los 10 resultados de los mejores individuos, se obtuvieron los valores de promedio, máximo y mínimo.

De acuerdo con los resultados de la función de aptitud en la Tabla 2, se puede observar que la mejor combinación de probabilidades es cuando ambos operadores de cruza y mutación tienen una probabilidad de 0.4.

De esta manera el resultado sugiere una calibración diferente a la que usualmente se encuentra en la literatura para una codificación binaria: una alta probabilidad de cruza y una baja probabilidad de mutación.

Para esta instancia de SC en particular, se requiere más exploración (i. e., valores de probabilidad de mutación más altos) en conjunto con una explotación moderada (valores de probabilidad de cruza más pequeños). El tamaño de la población y el número máximo de generaciones (condición de paro) fueron ajustados para mantener tiempos razonables de cada ejecución (alrededor de 90 minutos).

El AG fue ejecutado 10 veces para el segundo experimento con las probabilidades de cruza y mutación encontradas en el experimento previo (0.4 para ambas probabilidades de cruza y mutación). La tabla 3 muestra los resultados obtenidos. Cada fila de la tabla es una ejecución independiente.

Tabla 3. Experimento 2. Resultado de 10 ejecuciones del AG para reducir el número de características mientras se obtienen valores de clasificación competitivos.

Características (El mejor)	Precisión de clasificación (El mejor)	Aptitud (El mejor)
6	0.70	0.933
6	0.70	0.933
6	0.70	0.933
6	0.70	0.933
6	**0.76**	**0.947**
6	0.70	0.933
6	0.70	0.933
6	0.70	0.933
6	0.70	0.933
6	0.70	0.933

La solución encontrada para el AG propuesto tiene una reducción significativa del número de características (sólo 6), con una precisión de 0.76 y una aptitud de 0.947. Este resultado sugiere que existe información útil para la clasificación de reseñas en español sólo en un número reducido de características.

Finalmente, el tercer experimento compara el desempeño de la RNA usando el número total de características, 768, contra el mejor resultado, 6 en este caso, encontrado por el AG. La RNA usada para este experimento tiene la misma arquitectura que la usada previamente, además, la especificación del optimizador, la función de activación en la capa de salida y la función de pérdida, son los mismos que fueron usados en la función de aptitud del AG.

Con el objetivo de comparar ambas configuraciones de la red neuronal, 30 ejecuciones independientes fueron realizadas, respectivamente. Cada ejecución usa 70 reseñas para el entrenamiento y 30 para las pruebas de la red neuronal. Éstas generan 30 valores de precisión por cada configuración. La prueba estadística de Wilcoxon rank-sum, con un 95 % de confianza, fue ejecutada usando dichas precisiones de clasificación de las configuraciones.

Si el resultado del p-value es menor que 0.05, implica que los resultados tienen una diferencia significativa. De otra manera, las muestras no tienen una diferencia significativa. Los resultados se resumen en la Tabla 4.

El desempeño obtenido de la RNA usando el conjunto de características reducido por el AG, es mejor cuando se compara para la misma red pero con el conjunto de características original. Además, el número de operaciones de la red se reduce en consecuencia, y la reducción del tiempo de entrenamiento también es importante en comparación cuando se usan todas las características de representaciones vectoriales generadas por BETO. Finalmente, los resultados de las pruebas estadísticas indican una diferencia significativa entre los desempeños de ambas redes con un p-value de 1.2953e-06.

6. Conclusiones y trabajo futuro

En este artículo, se propuso una SC basado en envoltura con un AG para reducir el número de características de las representaciones vectoriales que provienen de la versión en español de BERT, BETO.

Tabla 4. Experimento 3. Resultados de la comparación de ambas configuraciones de la RNA con el conjunto de características original y el conjunto reducido.

Características	**Precisión (Min, Avg, Max)**			**Longitud de los datos**	**Operaciones de la red (incluye bias)**	**Tiempo de ejecución aproximado (Min, Avg, Max) en minutos**			**Wilcoxon p-value**
768	0.43	0.52	0.66	3,840,000	1,474,713,604	244	429	755	1.2953e-06
6	0.70	**0.69**	0.70	30,000	91,204	73	75	76	

El AG logró un desempeño muy competitivo, donde el número de características original fue reducido aproximadamente en un 99 %. Además, este número de características mejora la precisión de clasificación de la RNA en comparación de cuando se están usando las representaciones vectoriales con el número original de características.

El enfoque propuesto presenta una opción viable en el idioma español para generar modelos menos complejos, tratar con aquellas características de las representaciones vectoriales generadas por BETO, y de esta manera obtener una clasificación adecuada. De la misma forma, el costo computacional puede ser reducido en el proceso de entrenamiento. Sin embargo, es necesario mencionar que el costo computacional en el proceso de SC debe de ser considerado y reducido (ver el trabajo futuro al final de esta sección).

El trabajo futuro incluye: (1) considerar la implementación de la paralelización del AG para reducir el tiempo requerido en la función de aptitud, (2) aunado a lo anterior, se abre la posibilidad de utilizar un conjunto de datos nativo del español como el mencionado en el estudio [8], donde se propone una base de datos que contiene tres clases.

Si el número de clases aumenta, la complejidad de la RNA también lo hace. (3) Se puede incrementar el número de reseñas y tokens utilizados en el proceso de entrenamiento de cada individuo. (4) Por último, se propone utilizar clasificadores tradicionales de aprendizaje automático para comparar el desempeño contra la RNA presentada en este trabajo.

Agradecimientos. Los dos primeros autores agradecen el apoyo del Consejo Nacional de Ciencia y Tecnología (CONACyT), mediante una beca para realizar estudios de posgrado en la Universidad Veracruzana.

Referencias

1. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186 (2019) doi: 10.18653/v1/n19-1423
2. Fernandez, L. D.: IMDB dataset of 50k movie reviews (spanish) https://www.kaggle.com/datasets/luisdiegofv97/imdb-dataset-of-50k-movie-reviews-spanish
3. Gnana-Singh, D. A. A., Leavline, E. J., Priyanka, R., Priya, P. P.: Dimensionality reduction using genetic algorithm for improving accuracy in medical diagnosis. International Journal

of Intelligent Systems and Applications, vol. 8, no. 1, pp. 67–73 (2016) doi: 10.5815/ijisa.2016.01.08
4. Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Gonzalez-Agirre, A., Armentano-Oller, C., Rodriguez-Penagos, C., Villegas, M.: MarIA: Spanish language models. Procesamiento del Lenguaje Natural, pp. 39–60 (2021) doi: 10.26342/2022-68-3
5. Hernández-Hernández, J. C., Mezura-Montes, E., Hoyos-Rivera, G. J., Rodríguez-López, O.: Neuroevolution for sentiment analysis in tweets written in mexican spanish. pp. 101–110 (2021) doi: 10.1007/978-3-030-77004-4_10
6. Khaire, U. M., Dhanalakshmi, R.: Stability of feature selection algorithm: A review. Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 4, pp. 1060–1073 (2022) doi: 10.1016/j.jksuci.2019.06.012
7. Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., Potts, C.: Learning word vectors for sentiment analysis. pp. 142–150 (2011)
8. Pérez, J., Recart, E., Alves-Salgueiro, T., Furman, D., Fernández-Larrosa, P. N.: A spanish dataset for targeted sentiment analysis of political headlines. Electronic Journal of SADIO, vol. 22, no. 1, pp. 53–66 (2022)
9. Tian, W., Li, J., Li, H.: A method of feature selection based on Word2Vec in text categorization. In: 2018 37th Chinese Control Conference (CCC), pp. 9452–9455 (2018) doi: 10.23919/chicc.2018.8483345
10. Uğuz, H.: A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems, vol. 24, no. 7, pp. 1024–1032 (2011) doi: 10.1016/j.knosys.2011.04.014
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems, vol. 30, pp. 5999–6009 (2017)
12. Wang, K., Huang, J., Liu, Y., Cao, B., Fan, J.: Combining feature selection methods with BERT: An in-depth experimental study of long text classification. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 349, pp. 567–582 (2021) doi: 10.1007/978-3-030-67537-0_34
13. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. vol. 2, pp. 1–23 (2016) doi: 10.48550/arXiv.1609.08144
14. Xue, B., Zhang, M., Browne, W. N., Yao, X.: A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation, vol. 20, no. 4, pp. 606–626 (2016) doi: 10.1109/TEVC.2015.2504420