Agrupamiento de poemas de autores suicidas y no suicidas usando K-means y enjambre de partículas

Clustering of poems by suicidal and not suicidal authors using K-means and particle swarm optimization

Jairo Egberto Powell González
Benemérita Universidad Autónoma de Puebla, Puebla, México
jairo.powell@alumno.buap.mx
ORCID: 0000-0001-7113-3159

Maya Carrillo Ruiz
Benemérita Universidad Autónoma de Puebla, Puebla, México
maya.carrillo@correo.buap.mx
ORCID: 0000-0001-6152-456X

María Josefa Somodevilla García
Benemérita Universidad Autónoma de Puebla, Puebla, México
maria.somodevilla@correo.buap.mx
ORCID: 0000-0002-1972-2252

doi: https://doi.org/10.36825/RITI.09.18.002

Recibido: Diciembre 01, 2020
Aceptado: Marzo 05, 2021

Resumen: El suicidio es considerado un problema de salud pública, y su detección y tratamiento de manera temprana pueden contribuir a prevenirlo. La detección automática de indicadores de ideación suicida en textos tiene posibilidad de ser una herramienta útil para su prevención. En este trabajo, se reunió un corpus formado por poemas de doce poetas distintos, de los cuales seis cometieron suicidio. Se experimentó con dos representaciones vectoriales, una por número total de palabras y otra por palabras relacionadas con conceptos emocionales negativos. Los vectores se agruparon utilizando dos algoritmos: *K-means* y un híbrido de *K-means* con *Optimización por Enjambre de Partículas*. Se comparó la eficiencia de las representaciones vectoriales y de los algoritmos usados y se obtuvo que, por medio del algoritmo híbrido y del vocabulario relacionado con conceptos emocionales negativos, los grupos de poetas con ideación suicidada y sin ella pudieron ser distinguidos con una exactitud de hasta 0.98.

Palabras clave: *Agrupamiento, Metaheurísticas, K-Means, Optimización por Enjambre de Partículas, Detección de Ideación Suicida.*

Abstract: Suicide is considered a public health issue, and its early detection and treatment may contribute to its prevention. Automatic detection of suicidal ideation indicators within texts can be a useful tool to prevent it. In this work a corpus was compiled, which consists of poems written by twelve different poets, where six of them committed suicide. Two vector representations were experimented on, one with the total number of words and another with words related to negative emotional concepts. The vectors were clustered using two algorithms: *K-*

Means and a *K-Means* with *Particle Swarm Optimization* hybrid. The efficiency of the vector representations and the used algorithms were compared, obtaining as result that, through the hybrid algorithm and the negative emotional concepts vocabulary, the groups of poets with suicidal ideation and without it could be distinguished with an accuracy of 0.98.

Keywords: *Clustering, Metaheuristics, K-Means, Particle Swarm Optimization, Suicidal Ideation Detection.*

1. Introducción

De acuerdo con el Instituto Nacional de Estadística, Geografía e Informática [1] el suicidio ocupa el lugar 22 en la lista de causas principales de muerte en el total de la población. En 2017 representó un 0.9% del total de muertes y es considerado un problema de salud pública. La detección y tratamiento del suicidio de manera temprana pueden ayudar a su prevención. Tener un monitoreo del problema y de las conductas que llevan a él, permite llevar a estrategias que lo prevengan.

Como se muestra en el artículo de Ji *et al*. [2], considerando el aumento de acceso a Internet de las personas, la detección automática de indicadores de ideación suicida en textos sería una herramienta útil en la prevención de casos de suicidio. Las comunidades de Internet permiten a las personas expresar diversas emociones y pensamientos, y entre ellas algunas pueden indicar deseos de suicidio; por esto se vuelve importante el desarrollo de herramientas automáticas de detección en los entornos en línea. Normalmente los expertos en psicología y psiquiatría se encargan de su detección por medio de cuestionarios y estudios clínicos, pero se ha estudiado la posibilidad de apoyar en esta tarea utilizando técnicas de Aprendizaje Automático e Inteligencia Artificial para predecir la posibilidad de suicidio.

En este trabajo se realizó una recopilación propia de poemas, que están divididos en dos categorías: textos de autores que cometieron suicidio y autores que no lo cometieron. Se utilizaron dos algoritmos de agrupamiento conocidos sobre los textos, con distintos modos de extraer sus características en forma de vectores, con el objetivo de encontrar características que distingan los textos de los dos grupos de autores. Este trabajo se encuentra organizado de la siguiente manera: En la sección 2 del artículo se presenta el estado del arte en clasificación de textos con indicadores suicidas, en la tercera sección se especifican los métodos utilizados para extracción de características y agrupamiento. En la sección 4 se muestran y se analizan los resultados de los experimentos, y en la quinta sección se discuten las conclusiones y el trabajo futuro.

2. Estado del arte

Existen trabajos previos en los que se han aplicado técnicas de Procesamiento de Lenguaje Natural y Aprendizaje de Máquinas sobre textos para la detección de ideación suicida de forma automatizada. En el artículo de Pestian, Nasrallah, Matykiewicz, Bennett y Leenaars [3], se comparó el desempeño de algoritmos de aprendizaje y clasificación con el de profesionales de salud mental, en la tarea de distinguir notas suicidas reales de notas ficticias; encontraron que los profesionales clasificaban correctamente 63% de las notas, mientras que el mejor algoritmo lograba una exactitud de 78%.

Mulholland y Quinn [4] clasifican canciones compuestas por autores que cometieron suicidio de autores que no lo cometieron, obteniendo una precisión en la clasificación del 70%.

Zhang y Gao [5] utilizaron modelos de Procesamiento de Lenguaje para el análisis de similitud entre obras de poetas, por medio de vectores de palabras características. En su trabajo compararon los resultados de agrupamiento con el análisis literario e histórico existente para comprobar su confiabilidad; concluyeron que el agrupamiento de los autores se organiza en forma similar a las comparaciones realizadas por los analistas literarios.

Los algoritmos de agrupamiento tienen como objetivo la asignación de vectores multidimensionales de datos a un conjunto de grupos o clústeres. Estos métodos permiten identificar características naturales en los datos, de modo que se asocien en grupos con características similares; y la diferencia que tienen respecto a los algoritmos de clasificación, yace en que el proceso no está guiado por conocimiento previo de las clases a las que pertenecen los datos. El agrupamiento puede servir como un paso de preprocesamiento de datos en la construcción de modelos predictivos, donde se usa para asignar automáticamente etiquetas de clase a los datos, evitando la necesidad de que los expertos hagan esa asignación manualmente.

K-means es uno de los algoritmos más conocidos y utilizados en el agrupamiento de datos en espacios con dimensiones múltiples. En este método, un conjunto de vectores de dimensión *n* representan datos en un problema real de clasificación; éstos se asociaran en grupos de características similares, con el objetivo de que los miembros tengan mayor similitud con los demás de su grupo, que con los de otros grupos. El criterio de *K-Means* para determinar la similitud entre dos vectores es la distancia Euclidiana entre ellos [6]. El proceso de encontrar los grupos consiste en proponer un número de clases, que estarán asociadas a un punto en el espacio conocido como centroide. En forma iterativa, los puntos de datos se asociarán al centroide más cercano por su distancia, y estos últimos se moverán en el espacio hasta optimizar el agrupamiento.

Van derMerwe y Engelbrecht [7] exploraron el uso de *K-means* y un algoritmo conocido como *Optimización de Enjambre de Partículas* (Por sus siglas en inglés *PSO*) para solucionar problemas de agrupamiento de datos. *PSO* tiene como función solucionar problemas de optimización, realizando búsquedas en espacios multidimensionales. Este método simula el comportamiento de parvadas de pájaros para explorar el espacio del problema, hasta obtener una solución óptima [8]. Los resultados de Van derMerwe y Engelbrecht indicaron que usar un método híbrido de *K-means* y *PSO* obtiene mejores agrupamientos que *K-means* o *PSO* de forma independiente.

En la presente investigación se recopiló un conjunto de datos propio, o corpus; éste está conformado por obras pertenecientes a doce poetas, donde la mitad de ellos cometieron suicidio. Los textos están organizados por autor y por su pertenencia al grupo de Suicidas o No Suicidas. Se exploró el uso de los algoritmos mencionados para identificar características similares entre poetas que cometieron suicidio, por medio de sus escritos.

3. Materiales y métodos

Se reunió un corpus formado por poemas de doce poetas distintos, de los cuales seis cometieron suicidio. De cada poeta se recopilaron 50 poemas. Alejandra Pizarnik, Alfonsina Storni, Jaime Torres Bodet, José Antonio Ramos Sucre, José Asunción Silva y Leopoldo Lugones conforman el grupo de poetas Suicidas; mientras que Amado Nervo, Humberto Garza, José Emilio Pacheco, Octavio Paz, Ramón López Velarde y Salvador Díaz Mirón forman el grupo de poetas No Suicidas. Este conjunto aporta datos para experimentar con algoritmos de detección de características pertenecientes a una persona con deseos suicidas, y se utilizó para los experimentos con los algoritmos *K-Means* y *PSO*.

La elección de atributos que caractericen los textos juega un papel primordial en las tareas de procesamiento de lenguaje natural. Se realizaron experimentos para establecer atributos que permitan el agrupamiento de textos como pertenecientes a autores suicidas o no suicidas.

Para la representación de los textos se utilizó el modelo vectorial. La idea fundamental del modelo es que una colección de documentos se representa como un conjunto de vectores multidimensionales. El espacio vectorial sobre el que se definen estos vectores está generado por el conjunto de vectores de términos $\{t_i\}$ ($i = 1,\ldots, n$). Así, un documento d_j estará representado como la sumatoria de los vectores t_i multiplicados por una ponderación $w_{i,j}$, como se muestra en la Ecuación 1.

$$\overrightarrow{d_j} = \sum_{k=1}^{n} w_{i,j}\overrightarrow{t_i} \qquad (1)$$

Donde *n* es el número de palabras diferentes de la colección conocido como vocabulario. Así, los poemas se representaron por un vector de tamaño igual al vocabulario, utilizando como esquema de ponderación la frecuencia de términos (*tf*). Esta representación vectorial de los poemas se utilizó para los experimentos de agrupamiento.

Para la obtención del vocabulario, se experimentó con dos aproximaciones. En la primera aproximación, se extraen todas las palabras del conjunto completo de textos y se cuenta la frecuencia con que aparecen en textos individuales. Se hicieron tres formas de preprocesamiento de este vocabulario, obteniendo: Un conjunto con todas las palabras, es decir, sin procesar; un conjunto con todas las palabras vacías eliminadas (palabras comunes como artículos y preposiciones); y un conjunto donde se eliminan las palabras vacías y se emplea truncamiento. El truncamiento reduce palabras derivadas a formas más simples al eliminar partes como sufijos, conjugaciones y plurales, con el objetivo de reducir el tamaño del vocabulario. Los vectores que representan los textos se determinaron por medio de la frecuencia que tiene cada término del vocabulario dentro de ellos.

En la segunda aproximación se utilizó un conjunto de 134 palabras relacionadas con estados emocionales negativos, que se muestran en la Tabla 1.

Tabla 1. Lista de conceptos negativos.

Concepto	Palabras
Muerte	muerte, muertes, morir, muriendo, muero, mueren, muere, mueres, morimos, moría, morían, morías, morí, murió, morimos, murieron, mortal, moriría, morirían, morirías, moriré, morirán, morirás, morirá, mortales, murió.
Soledad	solo, sola, solos, solas, soledad, aislado, aislada, aislamiento, aísla, aíslan, aislados, aisladas, abandona, abandonan, abandonas, abandono, abandonado, abandonada, abandonaría, abandonarían, abandoné, abandonó, abandonaron, abandonados, abandonadas, abandonaré, abandonará, abandonarán, desolación, desolado, desolada, desolados, desoladas.
Tristeza	amargo, amarga, amargado, amargada, amargura, amargos, amargas, culpa, culpable, culpables, arrepentimiento, arrepentido, arrepentidos, arrepentida, arrepentidas, débil, débiles, debilidad, vacío, vacíos, vacía, vacían, vació, vaciaron, vaciaría, desesperado, desesperados, desesperante, desesperada, desespera, triste, tristes, tristeza, infeliz, infelicidad, infelices, miseria, miserable, miserables, desamparo, desamparado, desampara, desamparada, desamparados, desamparadas, desamparan, desesperanza, melancolía, melancólico, melancólicos, melancólica, melancólicas, deprimente, depresivo, depresión, aprensión.
Metáforas	gris, grises, grisáceo, negro, negros, negrura, negra, negras, azul, azules, oscuro, oscuros, oscuridad, oscura, oscuras, dormir, duermo, duermen, durmiendo.

Fuente: Elaboración propia.

En teoría, ciertas palabras podrían ser utilizadas con mayor frecuencia en textos de personas con deseos suicidas, como se menciona en Pestian et al. [3], donde las notas suicidas incluyen categorías de conceptos referentes a relaciones, estados emocionales, estados cognitivos y situaciones específicas.

Se eligieron términos relacionados con tres conceptos relacionados con suicidio: Muerte, Soledad y Tristeza. Para cada una de estas categorías, se recopilaron palabras cuyo significado tuviera asociación con estos conceptos y se realizó un conteo de su frecuencia en los textos de autores suicidas y no suicidas. Obtuvimos que para cada categoría, individualmente, sus palabras tienen mayor número de apariciones en textos de poetas suicidas, por lo cual contar las apariciones en conjunto podría permitir la distinción de textos de autores suicidas por medio de los algoritmos de agrupamiento anteriormente mencionados. Un ejemplo de esta forma de usar listas de palabras se menciona en Mulholland y Quinn [4], donde se identificó ideación suicida en canciones a partir de clases semánticas como: muerte, depresión, amor, sensualidad, entre otras.

Se experimentó además con una cuarta categoría de palabras que llamamos Metáfora, que contiene términos que no están directamente relacionados con las categorías anteriormente mencionadas, pero se encontraron con mayor frecuencia en los poemas de los autores suicidas y tienen asociaciones metafóricas negativas. En el artículo de Zhang y Gao [5] se discutió la búsqueda de similitud entre obras poéticas de diferentes autores por medio del vocabulario que utilizan. Observaron que la representación vectorial de textos y los algoritmos de agrupamiento formaban asociaciones entre escritores que tenían una relación, desde el punto de vista literario, como pertenecer a una misma era o movimiento.

Los documentos se representaron como vectores de dos dimensiones. La primera dimensión corresponde a la frecuencia relativa total de palabras en el documento asociadas a las categorías en conjunto: Muerte, Tristeza y Soledad. La segunda dimensión está formada por la frecuencia relativa total de las palabras en el documento que tuvieran relación con la categoría Metáforas. La frecuencia relativa de cada documento *f* está calculada como: el número de palabras en el documento que pertenecen a las categorías, dividida entre el número total de palabras del documento. La ecuación de cálculo de la frecuencia relativa se observa en la ecuación 2, donde *w* corresponde al conjunto de palabras en un documento y *C,* al conjunto de palabras dentro de las categorías mencionadas.

$$f = \frac{|w : w \in C|}{|w|} \qquad (2)$$

Teniendo una representación vectorial de los poemas, se experimentó con el tamaño de los documentos de cada poeta. Cada autor tiene 50 poemas, y éstos se pueden unir para formar documentos de mayor tamaño. Los números con que se experimentó fueron: 1 poema por documento, 2 poemas por documento y 10 poemas por documento.

Tras la representación vectorial de los documentos, se utilizó el algoritmo *K-means* para observar la formación de grupos con poemas correspondientes a autores suicidas y no-suicidas. Para determinar la eficacia del agrupamiento, se utilizó una métrica externa con base en el conocimiento de las etiquetas de los poemas escritos por autores de ideación suicida y los que no la tienen. *K-means* se aplico en las cuatro representaciones vectoriales de los documentos. El algoritmo puede observarse en la Tabla 2.

Tabla 2. Algoritmo *K-Means.*

Entrada: Vectores de datos z_p, número de centroides N_c, dimensión N_d
Salida: Vectores de centroides m_j y vectores z_p agrupados en clústeres C_j

1. Inicializar aleatoriamente los N_c vectores de centroides.
2. Repetir hasta cumplir un criterio de paro:
   a. Para cada vector z_p asignar el centroide m_j más cercano por distancia Euclidiana:

$$d(z_p, m_j) = \sqrt{\sum_{k=1}^{N_d} (z_{pk} - m_{jk})^2}$$

   b. Calcular los centroides por la media de sus vectores z_p

$$m_j = \frac{1}{n_j} \sum_{\forall z_p \in C_j} z_p$$

Fuente: Van der Merwe y Engelbrecth [7].

En el segundo experimento se utilizó *PSO* junto con *K-Means*, con base en el artículo de Van der Merwe y Engelbrecht [7], donde utilizan un híbrido de estos dos algoritmos para mejorar los agrupamientos obtenidos. El algoritmo se muestra en la Tabla 3.

Tabla 3. Algoritmo de *Optimización por Enjambre de Partículas.*

Entrada: Vectores de datos z_p, número de partículas N_c, dimensión N_d, posiciones de las partículas x_i, número de iteraciones t_{max}, velocidades de las partículas v_i.
Salida: Vectores de centroides m_j y vectores z_p Agrupamiento óptimo $\hat{y}_k$

1. Inicializar aleatoriamente cada partícula con N_c vectores de centroides.
2. Para t_1 hasta t_{max}:
   a. Para cada partícula i:
      i. Para cada vector de datos z_p:
         1. Calcular la distancia Euclidiana $d(z_p, m_{ij})$ a todos los centroides de clúster C_{ij}.
         2. Asignar a z_p al clúster más cercano C_{ij}.
      ii. Calcular la función de fitness:

$$f(x_i(t+1)) = \frac{1}{N_c} \sum_{j=1}^{N_c} \left(\frac{1}{C_{ij}} \sum_{\forall z_p \in C_{ij}} d(z_p, m_{ij}) \right)$$

      iii. Actualizar las posiciones de las partículas y la mejor posición de cada partícula y_i.

$$y_i(t+1) =$$

      iv. Actualizar los centroides de clúster en cada partícula con las ecuaciones:

1. $v_{i,k}(t+1) = wv_{i,k}(t) + c_1\left(y_{i,k}(t) - x_{i,k}(t)\right) + c_2\left(\hat{y}_k(t) - x_{i,k}(t)\right)$
2. $x_i(t+1) = x_i(t) + v_i(t+1)$

Donde $\hat{y}_k$ representa la mejor posición de todas las partículas en el tiempo *t*; *w*, c_1 y c_2 son constantes ajustables.

Fuente: Van der Merwe y Engelbrecht [7].

El algoritmo híbrido consiste en agrupar los vectores de documentos por *K-Means* y posteriormente usar *PSO* para mejorar el agrupamiento. En la tabla 4 se observa este proceso.

Tabla 4. Algoritmo híbrido *K-Means/PSO.*

Entrada: Textos de poemas p_{ij} ($i \in 12$ autores, $j \in 50$ poemas), tamaño de documento *s*
Salida: Vectores de centroides m_k y vectores z_p Agrupamiento óptimo $\hat{y}_l$

1. Para cada poeta *i*:
   a. d_p =Unión de poemas p_{ij} en documentos de tamaño *s*
2. Para cada documento *p*:
   a. *t*=número de palabras en el documento
   b. f_1 = frecuencia total de palabras pertenecientes a categorías Muerte, Soledad, Tristeza, en el documento.
   c. f_2 = frecuencia total de palabras en categoría Metáforas en el documento.
   d. $fr_1 = f_1/t$
   e. $fr_1 = f_2/t$
   f. *Vector* z_p *= (fr1, fr2)*
3. *AgrupamientoK = K-means (Vectores* z_p*)*
4. *AgrupamientoPSO = PSO(Vectores* z_p*, Centroides de AgrupamientoK)*

Fuente: Elaboración propia.

La implementación de los algoritmos y métodos mencionados en esta sección se realizaron con el lenguaje de programación *Python* y las siguientes bibliotecas: *Natural Language Toolkit*, para el procesamiento de texto; *NumPy*, para el manejo y almacenamiento de matrices numéricas; y *Scikit-Learn*, para el agrupamiento de datos.

4. Resultados

Se conocen las categorías de los documentos, perteneciendo a poetas suicidas o no-suicidas y se sabe que el algoritmo *K-means* hace una clasificación no supervisada de los documentos por similitud; entonces la eficacia se evalúa por medio del número de vectores agrupados en la categoría correcta. Suponiendo un agrupamiento correcto, los documentos que se conocen como Suicidas (SU) o No-Suicidas (NS) estarán separados en dos grupos distintos. Si no es correcta, los grupos tendrán documentos de diferentes clases mezclados.

En la Tabla 5 se observan los tres tipos de preprocesamiento del vocabulario completo de la primera aproximación. Cada poema de los autores se representó por un vector. También se muestra el tamaño de los vectores, determinado por la cantidad de palabras del vocabulario. En esta aproximación se tuvieron un total de 600 vectores, correspondientes a 50 vectores de cada uno de los 12 poetas.

Tabla 5. Vocabularios generados.

Preprocesamiento	Vectores por poeta	Tamaño de vectores
Vocabulario sin preprocesamiento	50	15952
Vocabulario sin palabras vacías	50	15727
Vocabulario sin palabras vacías y con truncamiento	50	8176

Fuente: Elaboración propia.

En la Tabla 6 se muestran los resultados del agrupamiento obtenido con *K-means* y el tipo de preprocesamiento que se usó para obtener el vocabulario, así como el número de documentos que se agruparon como NS o SU y el tipo de autor al que pertenecen en realidad. Se observa que el tipo de vocabulario utilizado afecta mínimamente el agrupamiento. No hubo una diferencia en la forma de usar el vocabulario general entre los poetas suicidas y no suicidas, que pudiera usarse como atributo de separación entre los dos grupos.

La exactitud del agrupamiento se calcula como el número de vectores agrupados correctamente, dividido entre el número total de vectores. Con el vocabulario completo, se agruparon correctamente 6 documentos de poetas suicidas y 294 de poetas no suicidas, y se tienen 600 vectores en total; es decir, se obtuvo una exactitud de 0.5. Con los otros dos vocabularios se obtuvieron 0.49 y 0.52, respectivamente, es decir, no hay clasificación adecuada entre las dos categorías.

Tabla 6. Agrupamiento *K-means*.

Vocabulario completo			
	Documentos No Suicidas	Documentos Suicidas	Total
Grupo NS	294	294	588
Grupo SU	6	6	12

Vocabulario sin palabras vacías			
	Documentos No Suicidas	Documentos Suicidas	Total
Grupo NS	299	300	599
Grupo SU	1	0	1

Vocabulario sin palabras vacías y truncando.			
	Documentos No Suicidas	Documentos Suicidas	Total
Grupo NS	277	263	540
Grupo SU	23	37	60

Fuente: Elaboración propia.

Dados los resultados de la primera aproximación, se realizó una segunda con el vocabulario anteriormente mencionado. En esta aproximación se usaron vectores de dos dimensiones. La frecuencia de palabras en las categorías Muerte, Soledad, Tristeza se almacena en la primera dimensión; y la frecuencia de palabras en la categoría Metáforas en la segunda dimensión, por estar relacionadas en forma más indirecta que las otras tres categorías.

El número de documentos por poeta se varió para observar como afectaba al agrupamiento. Como se mencionó anteriormente, cada autor tiene 50 poemas, los documentos se formaron uniendo un número de poemas, y los vectores característicos están dados por la frecuencia relativa de las palabras. En la Tabla 7 se muestran los números de poemas unidos para formar documentos y el total de vectores que se agrupan, a partir de todos los autores. Las recopilaciones de cada autor tienen un tamaño medio de 8190 palabras.

Tabla 7. Número de poemas por vector y total de vectores

Poemas por vector	**Total de vectores**
1	600
2	300
5	120
10	60

Fuente: Elaboración propia.

En la Tabla 8 se observan los resultados de clasificar solamente por *K-means*, con los diferentes tamaños de documento.

Usando un poema por documento, se agruparon correctamente 300 documentos que pertenecen a poetas suicidas, en un grupo SU, y 87 documentos pertenecientes a autores no suicidas se agruparon como NS. Por otro lado, 213 documentos de autores no suicidas se agruparon incorrectamente como SU, y ningún documento suicida se agrupó como NS. Esto representa una exactitud de 0.645.

Conforme se aumentó el número de poemas por documento, la exactitud del agrupamiento mejoró: 0.72 para dos poemas por documento, 0.875 para cinco poemas, y 0.95 para diez. Se puede observar en las tablas que los agrupamientos unen en el mismo clúster a todos los poetas suicidas, y que los errores provienen de documentos de poetas no suicidas erróneamente agrupados como suicidas.

Tabla 8. Agrupamiento *K-means* por lista de conceptos.

Documentos de 1 poema			
	Documentos No Suicidas	Documentos Suicidas	Total
Grupo NS	87	0	87
Grupo SU	213	300	513

Documentos de 2 poemas			
	Documentos No Suicidas	Documentos Suicidas	Total
Grupo NS	65	0	65
Grupo SU	85	150	235

Documentos de 5 poemas			
	Documentos No Suicidas	Documentos Suicidas	Total
Grupo NS	45	0	45
Grupo SU	15	60	75

Documentos de 10 poemas			
	Documentos No Suicidas	Documentos Suicidas	Total
Grupo NS	27	0	27
Grupo SU	3	30	33

Fuente: Elaboración propia

Después del experimento con *K-Means*, se realizó el proceso con el algoritmo híbrido. En la Tabla 9 se pueden observar los resultados obtenidos por este método. Como se observa, existe mayor separación entre los grupos NS y SU. Con un poema por documento la exactitud mejora de 0.645 a 0.72; con dos poemas, de 0.72 a 0.86; usando cinco, mejora a 0.92; y con diez poemas se obtiene 0.98 de exactitud.

Tabla 9. Agrupamiento *K-means/PSO* por lista de conceptos.

Documentos de 1 poema			
	Documentos No Suicidas	Documentos Suicidas	Total
Grupo NS	130	0	130
Grupo SU	170	300	470

Documentos de 2 poemas			
	Documentos No Suicidas	Documentos Suicidas	Total
Grupo NS	109	0	109
Grupo SU	41	150	191

Documentos de 5 poemas			
	Documentos No Suicidas	Documentos Suicidas	Total
Grupo NS	50	0	50
Grupo SU	10	60	70

Documentos de 10 poemas			
	Documentos No Suicidas	Documentos Suicidas	Total
Grupo NS	29	0	29
Grupo SU	1	30	31

Fuente: Elaboración propia

Los resultados obtenidos permiten hacer estas observaciones: entre mayor sea el número de poemas usados por documento, más se diferencia el uso de los términos en las categorías mencionadas; y el algoritmo híbrido mejora la exactitud del agrupamiento. Usar la frecuencia de vocabulario completo no obtuvo un agrupamiento eficiente de los documentos de poetas, a diferencia de usar conceptos relacionados con estados emocionales negativos.

5. Conclusiones

Se realizó una recopilación propia de poemas que incluyen autores que cometieron suicidio. Se experimentó con diferentes formas de extraer características de los textos en forma vectorial, y se agruparon los vectores por medio de los algoritmos *K-Means* y *Optimización por Enjambre de Partículas.* De acuerdo con los experimentos realizados se comprobó que se pueden diferenciar los poetas suicidas usando la frecuencia de vocabulario relacionado con estados mentales negativos, además de que la precisión del agrupamiento aumenta cuando se tienen textos de mayor extensión.

En los casos en que no se obtiene un agrupamiento óptimo, es observable que los documentos de autores Suicidas se encuentran asociados en el mismo clúster, y que un número de los No Suicidas quedan como falsos positivos a ideación suicida, principalmente con textos de poca extensión. Estos falsos positivos se reducen a medida que aumenta el tamaño de los textos. Puede decirse que en los textos de autores Suicidas están presentes con mayor frecuencia las categorías de palabras seleccionada, pero se necesitan una extensión de texto mayor para distinguir con mayor exactitud los textos de autores No Suicidas.

Los resultados obtenidos hacen evidente la posibilidad de identificar automáticamente la existencia de ideación suicida a partir de textos. Algunas de las limitaciones de la presente investigación se encuentran en el tamaño de la muestra, dada la cantidad de información que se requiere para encontrar un agrupamiento óptimo. Se puede extender este trabajo utilizando más conceptos o diferentes categorías de palabras que puedan mejorar la distinción entre textos pertenecientes a personas con ideación suicida; por ejemplo, usando léxico para análisis de sentimiento como el que se define en Perez-Rosas, Banea, y Mihalcea [9]; o diferentes construcciones, como en Mulholland y Quinn [4], donde buscan características como oraciones en modo pasivo y uso de pronombres singulares, además de palabras con connotaciones negativas.

Igualmente es posible realizar experimentación futura con textos más diversos, dado que los textos empleados son poemas escritos por autores profesionales y serán diferentes de textos escritos por gente de diferentes profesiones, edades o niveles de educación. También usarse algoritmos de clasificación y agrupamiento, o incluso técnicas de aprendizaje automático o aprendizaje profundo, para categorizar los textos de personas con ideación suicida.

6. Referencias

[1] Instituto Nacional de Estadística, Geografía e Informática. (2019). *Estadísticas a propósito del día mundial para la prevención del suicidio*. Recuperado de: https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2019/suicidios2019_Nal.pdf

[2] Ji, S., Pan, S., Li, X., Cambria, E., Long, G., Huang, Z. (2021). Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8 (1), 214-226. doi: https://doi.org/10.1109/TCSS.2020.3021467

[3] Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., Leenaars, A. (2010). Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3, 19-28. doi: https://doi.org/10.4137/BII.S4706

[4] Mulholland, M., Quinn, J. (2013). Suicidal tendencies: The automatic classification of suicidal and non-suicidal lyricists using NLP. Trabajo presentado en *Proceedings of the 6th International Joint Conference on Natural Language Processing,* Nagoya, Japón. Recuperado de: https://www.aclweb.org/anthology/I13-1079.pdf

[5] Zhang, L., Gao, J. (2017). A comparative study to understanding about poetics based on natural language processing. *Open Journal of Modern Linguistics*, 7 (5), 229-237. doi: https://doi.org/10.4236/ojml.2017.75017

[6] Rebala, G., Ravi, A., Churiwala, S. (2019) An introduction to machine learning (1era. Ed.). Switzerland: Springer. doi: https://doi.org/10.1007/978-3-030-15729-6

[7] Van der Merwe, D. W., Engelbrecht, A. P. (2003). Data clustering using particle swarm optimization. Trabajo presentado en *Congress on Evolutionary Computation*, Canberra, ACT, Australia. doi: https://doi.org/10.1109/CEC.2003.1299577

[8] Kennedy, J., Eberhart, R. (1995). Particle swarm optimization. Trabajo presentado en *International Conference on Neural Networks,* Perth, WA, Australia. doi: https://doi.org/10.1109/ICNN.1995

[9] Perez-Rosas, V., Banea, C., Mihalcea, R. (2012). Learning Sentiment Lexicons in Spanish. Trabajo presentado en *8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey. Recuperado de: http://lrec-conf.org/proceedings/lrec2012/pdf/1081_Paper.pdf