Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Detección de Tendencias Suicidas Utilizando Procesamiento de Lenguaje Natural

Por

JOSÉ LUIS SÁNCHEZ ARENAS

Tesis sometida como requisito parcial para obtener el grado de:

Licenciatura en Ciencias de la Computación

Supervisada por:

Dra. Maya Carrillo Ruiz

Dr. Luis Enrique Colmenares Guillén

Diciembre 2017

A mi mama por todo el apoyo, por esos días de levantarse temprano para acompañarme a tomar el transporte a las 4 de la mañana, por animarme siempre que sentía que las cosas estaban mal.

"Gracias por estar a mi lado y creer en mi"

A mis hermanas que me soportaban cuando tenía la luz prendida por hacer mi tarea hasta tarde.

A mi sobrina Michel que siempre está ahí para hacerme reír y hacer más fácil los momentos difíciles.

Agradecimientos

A mi asesora de Tesis la Doctora Maya Carrillo Ruiz por sus consejos su tiempo, la disposición para conmigo y mi proyecto, su apoyo constante y acertados comentarios.

A la Facultad de Ciencias de la Computación, por sus instalaciones y las facilidades proporcionadas durante mi estancia.

Al Laboratorio de Redes por el espacio que me brindo para realizar mis actividades y conocer a buenos amigos y compañeros.

A mis compañeros de Licenciatura con los cuales pase momentos inolvidables dentro de la facultad y con los cuales pasamos noches de desvelo realizando proyectos para pasar nuestras materias.

A todos mis profesores los cuales siempre tuvieron la amabilidad y el tiempo para resolver dudas y ayudarme en mi formación.

Resumen

En México, más de la mitad de los suicidios son consumados por personas con trastornos depresivos y cerca de uno de cada cuatro casos de suicidios se asocia al alcoholismo. El suicido es un fenómeno global que sucede en todas las regiones del mundo y en el trascurso de la vida. Un gran porcentaje de estas muertes se concentra en adolecentes donde los principales factores son el factor económico y problemas de autoestima.

La evolución de las nuevas tecnologías y el uso de Internet se incrementan año tras año. El acceso a las redes sociales ya es la principal actividad online, por encima de otras actividades. Se ha determinado que la mayoría de internautas en la red son adolecentes entre 13 y 24 años, mismos que representa el 26% y 20% de la población y donde los suicidios se presentan más frecuentemente.

Ante lo planteado anteriormente, el objetivo de esta investigación es dar los primeros pasos para generar la creación de herramientas que permitan detectar tendencias suicidas utilizando textos colocados en redes sociales por jóvenes. Con este objetivo, dado que no se cuenta con textos escritos por suicidas en español, se creó un corpus de poetas que cometieron suicidio y poetas que no lo cometieron. Después utilizando métodos de procesamiento de lenguaje natural y aprendizaje automático se separaron los poetas que cometieron suicidio de los que no, obteniendo un porcentaje de clasificación correcta del 74% con el algoritmo de máquinas de soporte vectorial.

Contenido

Índice de tablas

Índice de figuras

Capítulo 1. Introducción

La combinación de varios factores son los que llevan a las personas a quitarse la vida. Algunos estudios mencionan que este comportamiento se ve diferenciado, no solo por el sexo si también por grupos de edad.

Entre los jóvenes existen factores de índole laboral y económica; y en adultos mayores, la ausencia de seguridad social y pobreza.

En México, más de la mitad de los suicidios son consumados por personas con trastornos depresivos y cerca de uno de cada cuatro casos de suicidios se asocia al alcoholismo.

El suicido es un fenómeno global que sucede en todas las regiones del mundo y en el trascurso de la vida. No obstante, entre los jóvenes entre 15 y 29 años se trata de la segunda causa de muerte, pues representa el 8.5% del total de muertes en este grupo de edad con una taza de 13.5 suicidios por cada 100 mil jóvenes. [1]

Cada suicidio es una tragedia que afecta a familias, comunidades y países y tiene efectos duraderos para los allegados del suicida. El suicidio se puede producir a cualquier edad, y en 2015 fue la segunda causa de defunción en el grupo de entre 15 a 29 años en todo el mundo. [17]

El suicidio es un grave problema de salud pública; no obstante, es prevenible mediante intervenciones oportunas, basadas en datos fidedignos y a menudo de bajo coste. Para que las respuestas nacionales sean eficaces se requiere una estrategia de prevención del suicidio multisectorial e integral. [17]

La evolución de las nuevas tecnologías y el uso de Internet se incrementan año tras año; de acuerdo con datos de la Asociación Mexicana de Internet (AMIPCI), el tiempo promedio diario de conexión a Internet es de 6 horas y 11 minutos, de las cuales el 85% del tiempo es utilizado para acceder a redes sociales. El acceso a las redes sociales es ya la principal actividad online, por encima de enviar/recibir correos, aunque siguen utilizándose principalmente para actividades de ocio (juegos, ver películas, comunicación con amigos y descarga de música, entre otras). [3]

Ante lo planteado anteriormente, en esta investigación se utilizaron técnicas tradicionales de procesamiento de lenguaje natural para identificar tendencias suicidas en textos en español de diversos poetas, como primer paso hacia la creación de herramientas que permitan detectar tendencias suicidas utilizando textos colocados en redes sociales.

1.1. Descripción del problema

Una persona con tendencias suicidas tiene ciertas características diferentes a las de una persona que no presenta esta tendencia. La importancia de estudiar el tema del suicidio en México, se debe al interés de sustentar políticas públicas orientadas a la disminución de este problema de salud. [1]

La palabra "suicidio" proviene de las palabras latinas sui (uno mismo) y Caedere (matar), término acuñado en 1642 por el médico y filósofo Thomas Browne para distinguir entre el homicidio de uno mismo y el cometido hacia otra persona. De acuerdo con la Organización Mundial de la Salud (OMS), el suicidio se define como un acto deliberadamente iniciado y realizado por una persona en pleno conocimiento o expectativa de su desenlace fatal

Algunos de los patrones observados en la conducta suicida son: depresión, bipolaridad y esquizofrenia, todo ellos considerados trastornos psiquiátricos. Otra característica del fenómeno de los suicidios es que se presentan mayores tasas de suicidio en hombres, en personas de bajos ingresos, en desempleados y en usuarios de alcohol y drogas. [12]

El presente trabajo no intenta describir cada una de estas conductas, sólo plantea la utilización el procesamiento de lenguaje natural para ayudar a identificar ciertos patrones de escritura que utilizan las personas con tendencia al suicidio. De manera concreta este trabajo pretende encontrar las características estilísticas que permitan la diferenciación de textos escritos por personas con tendencias suicidas de los textos de personas no suicidas.

1.2. Objetivos generales y específicos de la tesis

Ante la situación planteada en los párrafos anteriores, estamos interesados en saber si podemos identificar tendencias suicidas a partir de los escritos de una persona, estableciendo los siguientes objetivos para esta investigación.

a) General

Utilizar técnicas de procesamiento de lenguaje natural para identificar escritos que muestran tendencias de suicidio empleando el idioma español.

b) Específicos

- Crear un corpus con al menos veinte textos de un mínimo de seis autores que cometieron suicidio.
- Crear un corpus con el mismo número de documento de autores que no cometieron suicidio.
- Determina los atributos para caracterizar los escritos de los autores mencionados.
- Utilizar técnicas de aprendizaje automático para ver si es posible distinguir escritos con tendencias suicidas de escritos que no las tienen.

1.3. Estructura de la tesis

En el capítulo 2 se presentan conceptos básicos útiles para comprender esta tesis, los cuales incluyen nociones de: pre-procesamiento de texto, representación de documentos y clasificación. En el capítulo 3 se presenta el estado del arte. En el capítulo 4 se presenta la utilización del procesamiento de lenguaje natural para identificar tendencias suicidas y se habla de manera detallada de los procedimientos utilizados para la elaboración de esta investigación. En el capítulo 5 se presenta los experimentos del método propuesto y se discuten los resultados. Finalmente, en el capítulo 6 se presentan las conclusiones y el trabajo futuro.

Capítulo 2. Conceptos generales

En este capítulo revisaremos los conceptos necesarios para comprender el resto del documento.

2.1. Representación de documentos

Para que un texto pueda ser entendido por la computadora, este tiene que representarse utilizando algún modelo como: el vectorial, probabilístico, booleano, modelos de lenguaje, entre otros.

Previamente a la utilización de los modelos mencionados, los textos tienen que pasar por una serie de pasos que los transforman en elementos útiles para dichos modelos, estos pasos se conocen como pre-procesamiento.

El pre-procesamiento del texto es la representación de un texto en un documento que brinden información de uno o más aspectos de su significado. Esta serie de transformaciones persigue, además, la reducción del texto a algún tipo de forma canónica que facilite el establecimiento de correspondencias durante el posterior proceso de representación del texto. [11]

El pre-procesamiento del texto consiste en general en los siguientes pasos:

a) Conversión del texto a minúsculas
b) Eliminación de caracteres especiales como: 'º','°','ª','…','“','_','#','`','"','"', '!', '$', '%', '&', '(', ')', '*', '+', ',', '-', '.', '/', ':', ';', '<', '=', '>', '?', '@', '[', '\\', ']', '^', '_', '{', '}', '~', '¿', '¡', '0', '1', '2', '3', '4', '5', '6', '7', '8', '9'.
c) Eliminación de palabras vacías.
   Una palabra vacía es una palabra muy común que no parece ayudar en la diferenciación de los textos, debido a su excesiva frecuencia, ya que esta anula su capacidad discriminante. Algunas de las características que nos pueden ayudar a detectar una palabra vacía son: tiene poca semántica (artículos, determinantes, pronombres, preposiciones, …), como son palabras muy frecuentes (más del 80%) y poco útiles, ahorramos espacio si las eliminamos. La eliminación de palabras vacías, depende mucho del: idioma y contexto.

d) Segmentación y Lematización

La segmentación y la lematización son dos métodos utilizados para reducir el tamaño del vocabulario. El vocabulario es el conjunto de palabras diferentes de una colección.

La lematización consiste en la reducción de una palabra a su raíz. La lematización utiliza el análisis morfológico de las palabras con el objetivo de eliminar las terminaciones flexivas y devolver la base de una palabra, que se conoce como lema. Por ejemplo:

Para fuimos sería ir y para am to be

La lematización cumple la función de hallar el lema correspondiente el cual es aceptado como el representante de todas las palabras flexionadas la cual es la alteración que sufren las palabras para expresar sus distintas funciones dentro de una oración.

Por otra parte, la segmentación busca la representación de las palabras en su forma más simple, empleando el truncamiento de palabras, la eliminación de afijos (prefijos, infijos y sufijos). Un ejemplo de esto es la reducción del siguiente conjunto de palabras: biblioteca, bibliotecario, bibliotecarios, bibliotecaria, bibliotecarias, biblioteconomía, a la palabra biblioteca.

Entre los algoritmos clásicos de segmentación destacan el de Porter Stemmer el cual es considerado el más popular, otro de ellos es el algoritmo de Lovins. En ambos casos podemos diferenciar dos fases: una fase de eliminación de sufijos en base a una lista prefijada de los mismos y una fase de recodificación de la cadena resultante de acuerdo a una serie de reglas.

Una vez realizada la parte de procesamiento, un punto importante para clasificar los documentos de una colección, es representarlos. A continuación, se explica el modelo vectorial que fue utilizado en este proyecto.

2.2. Modelo vectorial

En el modelo vectorial los documentos son representados como vectores de dimensión *t* (tamaño del vocabulario) y la similitud entre documentos se calcula como el coseno del ángulo entre los vectores que los representan. El modelo vectorial pondera los términos asignando pesos. La asignación de pesado es el proceso que tiene como finalidad conocer la importancia de los términos para representar un documento y permitir su posterior clasificación. Esto implica que se debe determinar el valor discriminativo de los términos de la colección, o lo que es lo mismo, la capacidad de los términos para diferenciar el contenido de los documentos en la colección.

Existen diferentes esquemas de ponderación de términos: el binario que únicamente considera la presencia o ausencia del término en el documento, asignando 1 o 0 respectivamente. El esquema basado en la frecuencia de aparición, para el cual el término que aparezca más veces en un documento será probablemente más importante que un término que lo haga solo una vez, un término que aparece en pocos documentos de la colección tendrá un mayor poder discriminador que los términos que aparecen en casi todos los documentos.

La frecuencia de aparición de un término o *tf*, es la suma de todas las ocurrencias o el número de veces que aparece un término en un documento. A este tipo de frecuencia de aparición también se le denomina "Frecuencia de aparición relativa" por que atañe a un documento en concreto y no a toda la colección. En la tabla 2.1 podemos ver la definición, características y finalidad de la ponderación *tf*. [15]

Ponderación *tf*	
Denominación	Frecuencia de aparición de término
Descripción	Número de veces que un término se repite en un documento, lo que permite determinar su capacidad de representación.
Finalidad	Representativa
Casos	Frecuencia de aparición *tf* baja. Representatividad elevada Frecuencia de aparición *tf* media. Frecuencia de aparición *tf* alta. Representatividad muy baja.

Tabla 2.1. Características de la ponderación *tf*

El cálculo de *tf* se efectúa una vez que el texto del documento ha sido pre-procesado, siguiendo los pasos mencionados en la sección anterior de acuerdo a la siguiente formula:

$$tf(n) = \sum_{i=0}^{D_i} n_i \qquad (2.1)$$

Dónde: *n* = término y *Di* = número de palabras en el documento.

Esta evaluación que sólo nos ofrece el número de veces que un término aparece en un documento se puede complementar con el factor *idf* (frecuencia inversa del documento para un término) el cual determina que la importancia de un término es inversamente proporcional al número de documentos en los que aparece dicho termino. Esto significa que cuanto menor sea la cantidad de documentos, así como

la frecuencia absoluta de aparición del término, mayor será su factor *idf* y a la inversa, cuanto mayor sea la frecuencia absoluta relativa por una alta presencia en todos los documentos de la colección, menor será su factor discriminatorio. La tabla 2.2 presenta una descripción de la ponderación *idf*. [15]

Ponderación *idf*	
Denominación	Frecuencia inversa del documento para un termino
Descripción	Es el coeficiente que determina la capacidad discriminatoria del término de un documento con respecto a la colección. Es decir, distinguir la homogeneidad o heterogeneidad del documento a través de sus términos.
Finalidad	Discriminatoria
Casos	Poder discriminatorio bajo. El término es genérico y aparece en la mayoría de los documentos. Poder discriminatorio medio. Poder discriminatorio alto. El término es especializado y aparece en pocos documentos.

Tabla 2.2. Características de la ponderación *idf*

El factor *idf* es único para cada término de la colección. El *idf* para un término dado *n* se calcula aplicando el logaritmo base 10 a la división del número total de documentos de una colección dividido entre el número de documentos de la colección en los que aparece el termino *n*. Al valor resultante se le suma 1 para evitar valores iguales a cero cuando los términos con *idf* son muy bajos [15]. La frecuencia inversa se calcula como fórmula:

$$IDF_{(n)} = log_{10} \frac{N}{DF_{(n)}} + 1 \tag{2.2}$$

Dónde: N es número total de documentos en la colección, $DF_{(n)}$es número de documentos en el que aparece el término *n*.

Otro esquema de pesado que permite capturar la importancia del término dentro del documento y dentro de la colección es la ponderación *tf-idf*.

La ponderación *tf-idf*, de un término *n* es el producto de su frecuencia de aparición *tf* en un documento *d* por su frecuencia inversa de documento (*idf*) tal como refleja la fórmula 2.3.

$$\text{tf\_}idf_{(n,d)} = tf_{(n,d)} * \ idf_n \tag{2.3}$$

Dónde**:** $tf_{(n,d)}$ es la frecuencia de aparición de un término *n* en un documento *d* e idf_n es la frecuencia inversa del documento del término *n*.

Una vez explicado cada uno de los esquemas de ponderación que son utilizados, a continuación, se explica lo que es la clasificación de textos, utilizada en este trabajo para categorizar documentos dentro de las categorías suicida y no suicida.

2.3. Clasificación

La clasificación se conoce también como aprendizaje automático, que busca lograr que un ordenador pueda solucionar por si sólo problemas que requieran ciertas habilidades más allá de la mera capacidad de cálculo. [14]

Un proceso de clasificación automática comienza con la recopilación y clasificación manual de un conjunto de datos (datos de entrenamiento), después se llevan los documentos a una representación adecuada para que finalmente se puedan aplicar distintos algoritmos de clasificación y así obtener la clasificación de los documentos.

Existen dos tipos de aprendizaje automático, el aprendizaje no supervisado y el aprendizaje supervisado. El objetivo del aprendizaje no supervisado es crear un método de análisis automático donde un modelo es ajustado a las observaciones y el cual no requiere un aprendizaje a priori, sino que trata a los objetos de entrada como variables aleatorias. Algunas de sus características son: solo requiere instancias, pero no etiquetas, sirve para entender y resumir datos, suele no requerir tiempo de entrenamiento a diferencia de los métodos supervisados.

Esté método de clasificación descubre en los datos de entrada y de forma autónoma: características, regularidades, correlaciones y categorías.

Por otra parte, el aprendizaje supervisado es capaz de crear una función la cual pueda predecir el valor correspondiente a cualquier objeto de entrada valido después de haber visto una serie de ejemplos, los datos de entrenamiento. Para aplicar un efectivo análisis supervisado tenemos que considerar varios pasos: determinar el tipo de ejemplos de entrenamiento, crear el conjunto de entrenamiento, el cual debe contener características propias, determinar la función de ingreso de la representación, determinar la técnica de aprendizaje y el ajuste de parámetros de aprendizaje, asignación de un etiquetado a cada uno de los documentos como suicida para los poemas de escritores suicidas y nosuicida para poemas de escritores no suicidas. A continuación, la figura 2.1 muestra un diagrama del proceso de clasificación supervisada en la cual, se parte de una serie de clases o categorías prediseñadas, en las cuales el proceso de calcular patrones en base a los objetos preclasificados se conoce como entrenamiento. [19]

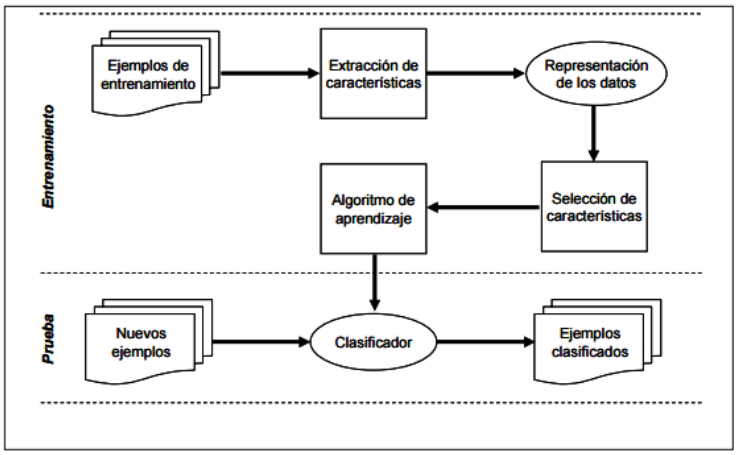

Figura 2.1 Aquí podemos observar el paradigma de aprendizaje supervisado, el cual intenta aprender conceptos a través de ejemplos de éstos. El clasificador construido utiliza los conceptos aprendidos para clasificar nuevos ejemplos fuente. [16]

Diversas pruebas se han realizado para comparar el rendimiento de los clasificadores sin embargo no hay un método de clasificación único que funcione para todos los problemas. Bajo esta perspectiva, se necesitan ciertas métricas comunes que permitan compararlos, tales como la precisión, recuerdo y medida F. [18]

La precisión se define como la cantidad de documentos clasificados y que son relevantes para nuestro sistema eso se ve representado en la fórmula:

$$Precision(P) = \frac{cantidad\_de\_docuementos\_relevantes\_recuperados}{cantidad\_de\_documentos\_recuperados}$$

(2.4)

El recuerdo se define como la cantidad de documentos relevantes clasificados y ayuda a la evaluación del sistema para encontrar todos los documentos relevantes la fórmula 2.5 muestra esta definición.

$$Recuerdo(R) = \frac{cantidad\_de\_docuementos\_relevantes\_recuperados}{cantidad\_de\_documentos\_relevantes}$$

(2.5)

Por último, la Medida F combina precisión y recuerdo en un único valor 0 o 1 donde el valor de F será alto cuando ambas componentes tengan valores altos.
Si F = 0 no se han clasificado documentos relevantes, mientras que si F=1 se han clasificado todos los documentos relevantes, la fórmula 2.6 representa esta definición. [18]

$$F = 2 * \frac{Precision * Recuerdo}{Precision + Recuerdo}$$

(2.6)

Algunos algoritmos de clasificación son las redes neuronales, las máquinas de soporte vectorial, el algoritmo de los k-vecinos más cercanos, el clasificador bayesiano ingenuo, los arboles de decisión. A continuación, se explican brevemente los utilizados en el presente trabajo.

2.4. Naive bayes

Uno de los algoritmos más utilizados para la clasificación de textos es el denominado Naive Bayes. Este clasificador es de tipo probabilístico, el cual se basa en el cálculo de distribuciones de probabilidad en función de datos observados. [16]

Naive Bayes construye modelos que predicen la probabilidad de posibles resultados. Este clasificador obtiene la probabilidad posterior de cada clase, C_i, (clase de clasificación) usando la regla de Bayes, como el producto de la probabilidad a priori de la clase o la probabilidad de todos los casos donde ocurra C_i por la probabilidad condicional de los nuevos datos de entrada E (documentos de entrada) o lo que es lo mismo el grado de presunción de que C_i dada la clase a

la cual puede pertenecer, dividido por la probabilidad de los datos. Esto puede representarse con la fórmula: 2.7. [18]

$$P(C_i|E) = \frac{P(C_i)P(E|C_i)}{P(E)} \tag{2.7}$$

Este clasificador asume que los atributos son independientes entre sí dada la clase, así que la probabilidad se puede obtener por el producto de las probabilidades condicionales individuales de cada atributo dado el nodo clase. Podemos observar esto en la siguiente formula: 2.8

$$P(C_i|E) = P(C_i)P(E_1|C_i)P(E_2|C_i) \dots P(E_n|C_i)C_i)/P(E) \tag{2.8}$$

2.5. Máquinas de soporte vectorial

Una máquina de soporte vectorial (SVM) es un modelo que representa a los puntos de muestra en el espacio, separando las clases en dos espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre dos puntos.

Las máquinas de soporte vectorial no se centran en construir hipótesis que cometan pocos errores, si no lo que pretenden es producir predicciones en las que se pueda tener mucha confianza, aun a costa de cometer ciertos errores. Las SVM buscan un hiperplano que estructuralmente tenga poco riesgo de cometer errores ante datos futuros.

Dado un conjunto linealmente separable de ejemplos de entrenamiento *S* (de muestras) formada por *n* ejemplos, es decir $s = \{(x_1,\ y_i), \dots (x_n,\ y_n)\}$, donde cada x_i pertenece al espacio de entrada X y la clase $y_i \in \{+1, -1\}$, existen muchos hiperplanos capaces de separar las clases aunque no todos son igual de buenos. Parece evidente, por tanto, que el hiperplano que este más alejado de los ejemplos de ambas clases, es decir, que defina una frontera más ancha es más resistente al ruido que puedan tener los ejemplos de entrenamiento y es menos probable que cometa errores ante datos futuros. Obsérvese la figura 2.2. [18]

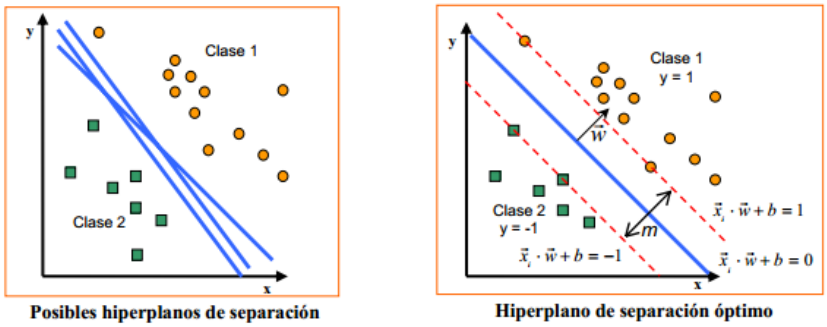

Figura 2.2 Ejemplo de hiperplanos para la separación de clases. Fuente. [18]

Esa separación entre el hiperplano y cada una de las dos clases es lo que se denomina margen, por tanto debe buscarse entre todos los hiperplanos validos aquel que presente un margen mayor, considerando como margen geométrico la distancia del hiperplano al ejemplo más próximo de cada clase. [14]

Capítulo 3. Estado del campo o del arte

Hablar sobre suicidio en la adolescencia resulta difícil por varias razones: porque se considera un tema trágico, un tabú que preferimos no mencionar, por temor a que se incremente el riesgo de los adolescentes a quitarse la vida; por considerar que el suicidio no es tan frecuente en estas edades, pues es una etapa de la vida con muchas potencialidades para desarrollar una existencia creativa; por creer que cuando ocurre, es un acto impulsivo, no bien pensado y por considerar que es una tarea exclusiva de psiquiatras, psicólogos y médicos, en el que no pueden intervenir otros sectores de la población [5].

J.P. Pestian, et al. en [6] presentan un trabajo en el que toman como referencia documentos que dejan las victimas antes de su acto suicida, y utilizan técnicas de procesamiento de lenguaje natural y clasificación automática de textos para detectar posibles conductas suicidas. De acuerdo con este articulo existen tres categorías en las que se clasifican los pacientes suicidas: aquellos que piensan en suicidarse, los que intentan suicidarse y los que se suicidan. Trabajan con un corpus obtenido de notas suicidas reales, y las mezclan con notas escritas por un grupo denominado simuladores al cual se le pide que escriban notas suicidas. Utilizan técnicas de procesamiento de lenguaje natural y máquinas de soporte vectorial (SVM), para identificar las notas escritas por suicidas reales, obtienen entre 60% y 79% de exactitud en la clasificación.

En los últimos años se han publicado varias investigaciones sobre esta temática [7, 8, 9, 10], algunos de estos publicados por Biomedical Informatics Insights, Libertas Academica, la cual edita 83 revistas internacionales de ámbito científico, técnico y médico. Varios de estos trabajos fueron presentados en el quinto congreso i2b2/ VA/Cincinnati Shared-Task, en la que participaron 106 científicos en un total de 24 equipos en desafíos de procesamiento de lenguaje natural para datos clínicos en Washington, DC, realizado el 21 de octubre de 2011. Los corpus utilizados en estas tareas fueron escritos por 1319 personas antes de morir por suicidio. Y fueron recogidas entre los años 1950 y 2011 por el Dr. Edwin Shneidman y Cincinnati Children’s Hospital Medical Center. [4]

Todas las investigaciones mencionadas, tienen en común que utilizan técnicas de clasificación y procesamiento de lenguaje natural para resolver la tarea.

El trabajo [7] propone un enfoque híbrido que utiliza reglas basadas en técnicas de Machine Learning. Para resolver el problema, se utiliza una clase basada en reglas

y un modelo SVM. Un conjunto de funciones sintácticas y semánticas son seleccionadas para cada frase y con ello construyen reglas para entrenar un clasificador.

Para esta tarea se logró obtener una precisión de 41.79%, un recuerdo de 55.03% y una medida F de 47.50%. El promedio general de la medida F de todas las presentaciones fue 48.75% con una desviación estándar de 7%. Teniendo en cuenta los errores comunes en los datos de entrenamiento, se preparó un script para hacer automáticamente los reemplazos necesarios. Sin embargo, para el resto de las palabras mal escritas, el sistema requirió que el usuario seleccionara manualmente las correcciones de la lista de sugerencias, en total se contó con 900 notas de las cuales 600 de ellas de usaron para entrenamiento y 300 notas se utilizaron para realizar pruebas. El desempeño del sistema se midió utilizando un promedio de tres mediciones estándar: recuerdo (R), precisión (P) y medida F (F). Se compara el rendimiento del sistema cuando se termina de aplicar métodos de aprendizaje basados en reglas o SVM con los experimentos en los que se aplicó una combinación de ambos. En un experimento, se aplicó el aprendizaje automático a un número limitado de emociones para las cuales el clasificador generó resultados aceptables. Entonces, para incorporar otras emociones, las reglas primero fueron aplicadas y los clasificadores fueron utilizados en 4 emociones con resultados aceptables de la clasificación. Usando el aprendizaje de la máquina sin reglas resultó en la media F de 41.96%, mientras que el uso de reglas solo resultó en la medida F de 45.95%. Luego se aplicó la combinación de aprendizaje automático y reglas para todas las emociones y la medida F aumentó a 47.36%. En el experimento donde se eliminaron las emociones con pequeñas instancias de entrenamiento (por ejemplo, "abuso") tuvimos un incremento del 0.14% en el desempeño y alcanzamos el 47.50% de la medida F.

Para una oración dada, los verdaderos positivos son el número de emociones que son asignados por el sistema y que existen en el estándar de oro. Los falsos positivos son el número de emociones que el sistema asigna pero que no existen en el estándar de oro. Las emociones que se asignan a la oración sólo en el estándar de oro, pero no por el sistema se consideran como falsos negativos.

Un trabajo más es [8] Esta tarea compartida se centró en el desarrollo de sistemas automáticos que identifican, a nivel de oración, el texto afectivo de 15 emociones específicas de notas de suicidio. Se propone un modelo híbrido que incorpora una serie de técnicas de procesamiento del lenguaje natural, incluyendo la localización de palabras clave basadas en el léxico, la identificación de señales de emoción basadas en CRF (Conditional Random Field) o campo aleatorio condicional en español y la clasificación de emociones basada en el aprendizaje automático. Los resultados generados por diferentes técnicas se integran utilizando diferentes estrategias de fusión basadas en el voto. El sistema automatizado se comportó bien contra el estándar de oro anotado manualmente y logró resultados alentadores con un puntaje promediado de 61.39% en el reconocimiento textual de

emociones, que se ubicó en el primer lugar de 24 equipos participantes en este desafío.

En el siguiente trabajo [9] los autores presentan un sistema desarrollado para el 2011 i2b2 Challenge on Sentiment Classification, cuyo objetivo era clasificar automáticamente las oraciones en notas de suicidio utilizando un esquema de 15 temas, en su mayoría emociones. El sistema combina el aprendizaje automático con una metodología basada en reglas. Las características utilizadas para representar un problema se basaron en las propiedades léxico-semánticas de las palabras individuales, además de las expresiones regulares utilizadas para representar los patrones de uso de la palabra a través de diferentes temas. Un clasificador naïve Bayes fue entrenado utilizando las características extraídas de los datos de formación consistente en 600 documentos escritos manualmente sobre suicidio. La clasificación se realizó entonces utilizando el clasificador naïve Bayes, así como un conjunto de normas de patrones de concordancia. El rendimiento de la clasificación se evaluó con un estándar de oro preparado manualmente que consta de 300 notas de suicidio, en el que 1.091 de un total de 2.037 oraciones se asociaron con un total de 1.272 anotaciones(revisar). Los sistemas que compiten se clasificaron usando la medida F micro-promediada como la métrica primaria de evaluación. El sistema alcanzó la medida F del 53% (con un 55% de precisión y un 52% de recuerdo), lo cual fue significativamente mejor que el promedio del 48.75% logrado por los 26 equipos participantes.

En el trabajo [10] se menciona que el suicidio es un importante problema de salud pública. En 2007, el suicidio fue la décima causa de muerte en Estados Unidos, con 34.598 muertes, con una tasa global de 11,3 muertes por 100.000 personas. La tasa de suicidios para los hombres es cuatro veces mayor que la de las mujeres, el comportamiento suicida es complejo, con riesgos y desencadenantes biológicos, psicológicos, sociales y ambientales. Algunos factores de riesgo varían con la edad, el género o el grupo étnico y pueden ocurrir en combinaciones o cambiar con el tiempo.

Los factores de riesgo de suicidio incluyen depresión, intentos previos de suicidio, antecedentes familiares de trastorno mental o abuso de sustancias, antecedentes familiares de suicidio, armas de fuego en el hogar y encarcelamiento. Los hombres y los ancianos son más propensos a tener intentos fatales(revisar) que las mujeres y jóvenes. Las notas de suicidio han sido estudiadas durante mucho tiempo como una forma de entender los motivos y pensamientos de quienes intentan o completan un esfuerzo suicida. Dado el impacto del suicidio y otros trastornos mentales, el objetivo general de los organizadores del congreso i2b2 fue desarrollar métodos para analizar el texto libre neuropsiquiátrico subjetivo. Para lograr ese objetivo, este desafío se centró en el análisis del sentimiento, predijo la presencia o ausencia de 15 emociones en notas de suicidio.

Los equipos exploraron múltiples enfoques que combinan reglas basadas en la expresión regulares, la minería de texto (STM) y un enfoque que aplica pesos al texto mientras se contabilizan varias etiquetas. En general, el mejor sistema alcanzó una puntuación de promedio de 0.5023, ligeramente por encima de la media de los 26 equipos que compitieron (0.4875).

Otro trabajo desarrollado para el análisis de textos suicidas es presentado por M. Mulholland y J. Quinn en [2]. Ellos analizan a 5 cantautores del idioma ingles que cometieron suicidio para generar un sistema de clasificación a partir de 63 canciones de cada autor. En este trabajo se presenta la creación de un corpus de canciones de suicidas y no suicidas de sexo masculino. El conjunto de entrenamiento se compone de 533 canciones, de las cuales 253 eran escritas por cuatro letristas que no cometieron suicidio y 280 por cinco letristas que lo cometieron. El conjunto de pruebas se compone de 63 canciones de 5 letristas no suicidas y 46 de canciones de 4 letristas suicidas. Los autores obtienen 70.6% de notas clasificadas correctamente de letristas que cometieron suicidio y los que no lo cometieron. [2]

Ninguno de los trabajos mencionado ha experimentado con textos en lenguaje español. El objetivo de este trabajo es generar métodos de clasificación para este idioma.

Capítulo 4. Método propuesto

En el siguiente capítulo se presenta el método propuesto, la manera de selección de los autores y las condiciones que cumplen para aparecer en este trabajo, cuantos documentos fueron recolectados para cada caso (autores suicidas y autores no suicidas) y un ejemplo de los documentos obtenidos.

4.1. Creación de corpus

Se recopiló un corpus de 600 documentos en el que se presentan poemas de doce poetas de los cuales seis cometieron suicidio y seis no lo cometieron. Se utilizaron dos condiciones básicas para el armado de este corpus las cuales son:

1) Autores de la misma época
2) Autores del mismo genero

Aunque en algunos casos no se cumplen ambas condiciones se trató que alguna de las dos fuese válida, esto con el fin de lograr consistencia entre los poemas del corpus.

La tabla 4.1 lista el nombre de los poetas considerados para la clasificación, se listan de acuerdo a su tipo de muerte (suicidio, no suicidio).

Suicida	**No suicida**
ALEJANDRA PIZARNIK	AMADO NERVO
ALFONSINA STORNI	HUMBERTO GARZA
JAIME TORRES BODET	JOSÉ EMILIO PACHECO
JOSÉ ANTONIO RAMOS SUCRE	OCTAVIO PAZ
JOSÉ ASUNCIÓN SILVA	RAMÓN LÓPEZ VELARDE
LEOPOLDO LUGONES	SALVADOR DÍAZ MIRÓN

Tabla 4.1 Autores de los poemas del corpus

El corpus fue recopilado de la web, y costa de 600 poemas, 300 de autores que cometieron suicidio y 300 de poetas que murieron de forma natural, todos estos escritos en español. Las tablas 4.2 y 4.3 presentan algunas características de los autores, así como el número de documentos considerados en el corpus para cada uno de ellos.

Poeta	Número de documentos	Genero	Tipo
ALEJANDRA PIZARNIK	50	Poesía	Suicida
ALFONSINA STORNI	50	Poesía	Suicida
JAIME TORRES BODET	50	Poesía	Suicida
JOSÉ ANTONIO RAMOS SUCRE	50	Poesía	Suicida
JOSÉ ASUNCIÓN SILVA	50	Épica, Lirica	Suicida
LEOPOLDO LUGONES	50	Poesía	Suicida

Tabla 4.2 Detalles sobre los poemas recuperados del tipo suicida

Poeta	Numero de Documentos	Genero	Tipo
AMADO NERVO	50	Novela, Poesía, ensayo	No suicida
HUMBERTO GARZA	50	Poesía	No suicida
JOSÉ EMILIO PACHECO	50	Poesía, Cuento, Novela	No suicida
OCTAVIO PAZ	50	Poesía, Ensayo	No suicida
RAMÓN LÓPEZ VELARDE	50	Poesía	No suicida
SALVADOR DÍAZ MIRÓN	50	Poesía	No suicida

Tabla 4.3 Detalles sobre los poemas recuperados del tipo no suicida

Como se mencionó anteriormente se cuidó que se tratara de poetas del mismo género para tener documentos homogéneos que ayudarán durante el proceso de clasificación.

Para ilustrar el contenido de los poemas, se comparan los fragmentos de dos poemas de cada tipo, las tablas 4.4 y 4.5 representan este proceso.

Poetas Suicidas

Alejandra Pizarnik	**Alfonsina Storni**
A la espera de la oscuridad Ese instante que no se olvida Tan vacío devuelto por las sombras Tan vacío rechazado por los relojes Ese pobre instante adoptado por mi ternura Desnudo de sangre de alas Sin ojos para recordar angustias de antaño Sin labios para recoger el zumo de las violencias perdidas en el canto de los helados campanarios.	¡Adiós! Las cosas que mueren jamás resucitan, las cosas que mueren no tornan jamás. ¡Se quiebran los vasos y el vidrio que queda es polvo por siempre y por siempre será! Cuando los capullos caen de la rama dos veces seguidas no florecerán... ¡Las flores tronchadas por el viento impío se agotan por siempre, por siempre jamás! ¡Los días que fueron, los días perdidos, los días inertes ya no volverán! ¡Qué tristes las horas que se desgranaron bajo el aletazo de la soledad!

Tabla 4.4 La tabla presenta un ejemplo de dos poetas suicidas diferentes

Poetas No suicidas

Amado Nervo	**Humberto Garza**
¡Está bien! Porque contemplo aún albas radiosas y hay rosas, muchas rosas, muchas rosas en que tiembla el lucero de Belén, y hay rosas, muchas rosas, muchas rosas Gracias, ¡está bien!	Extraterrestre Como una viajera interplanetaria que no comprendía los gestos de alegría o de enojo, así eras tú. Con tus ojos mágicos y extraños me veías llorar y golpear la tierra, me veías rechinar los dientes, en momentos raspados por higueras que tirarían sus hojas en noviembre.

Tabla 4.5 En la tabla se presentan dos ejemplos de poemas de escritores no suicidas

En los poemas de las tablas anteriores se pueden observar algunas diferencias entre estos, aun sin aplicar ningún procesamiento computacional. En los poemas del tipo no suicida podemos observar expresiones de alegría, palabras que muestran un estado de ánimo feliz, contento, se muestran palabras que representan amor a la vida, a la naturaleza. Sin en cambio en los ejemplos de poemas del tipo suicida se observa un estado de animo de tristeza y desolación además de hablar de cosas que incitan a la muerte, hablan sobre obscuridad y despedida es desde ese momento en el que se puede observar que existe un comportamiento diferente en las personas que escribieron esos textos.

En el capítulo 2 se presentaron los conceptos generales de pre-procesamiento de documentos con aprendizaje supervisado, en este capítulo se aplicarán estos conceptos. En el diagrama de bloques 4.1 se resume el proceso.

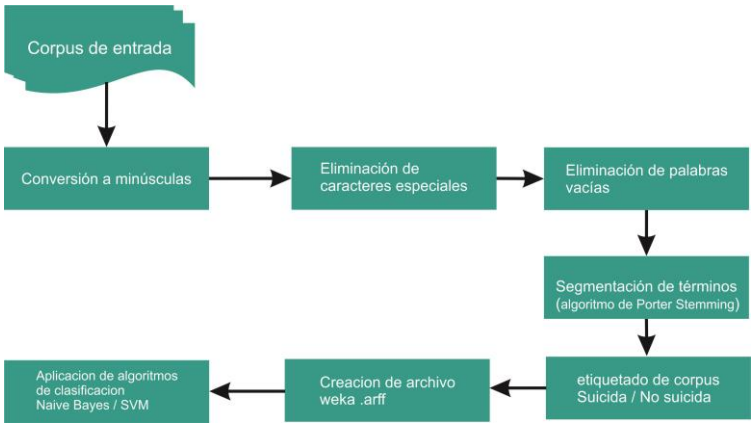

Figura 4.1. Diagrama de bloques que describe el proceso propuesto.

4.2. Representación de los poemas

Ahora se presentan los pasos del pre-procesamiento por los que pasando los poemas y los cambios que fueron adquiriendo:

a) Conversión a minúsculas, como se explicó en el capítulo 2 y en el diagrama 4.1 todas las palabras deben homogeneizarse y un paso para ello fue la conversión de letras mayúsculas a minúsculas, la tabla 4.7 muestra el resultado de esta operación.

Texto de entrada	Texto convertido a minúsculas
¡Está bien! Porque contemplo aún albas radiosas y hay rosas, muchas rosas, muchas rosas en que tiembla el lucero de Belén, y hay rosas, muchas rosas, muchas rosas gracias, ¡está bien! Porque en las tardes, con sutil desmayo, piadosamente besa el sol mi sien, y aun la transfigura con su rayo: gracias, ¡está bien! Porque en las noches una voz me nombra (¡voz de quien yo me sé), y hay un edén escondido en los pliegues de mi sombra: gracias, ¡está bien Porque hasta el mal en mí don es del cielo, pues que, al minarme va, con rudo cielo, desmoronando mi prisión también; porque se acerca ya mi primer vuelo: gracias, ¡está bien!	está bien porque contemplo aún albas radiosas y hay rosas muchas rosas muchas rosas en que tiembla el lucero de belén y hay rosas muchas rosas muchas rosas gracias está bien porque en las tardes con sutil desmayo piadosamente besa el sol mi sien y aun la transfigura con su rayo gracias está bien porque en las noches una voz me nombra voz de quien yo me sé y hay un edén escondido en los pliegues de mi sombra gracias está bien porque hasta el mal en mí don es del cielo pues que al minarme va con rudo celo desmoronando mi prisión también porque se acerca ya mi primer vuelo gracias está bien

Tabla 4.6. Ejemplo del procesamiento conversión a minúsculas

b) Eliminación de caracteres especiales, en los textos de la tabla 4.7. se puede observar cómo se han eliminado estos caracteres.

A continuación, se muestra los caracteres especiales que fueron eliminados de los poemas:

non_words =list(['º','°','ª','ª','…','“','–','#','`','"',"", '!', '$', '%', '&', '(', ')', '*', '+', ',', '-', '.', '/', ':', ';', '<', '=', '>', '?', '@', '[', '\\', ']', '^', '_', '{', '}', '~', '¿', '¡', '0', '1', '2', '3', '4', '5', '6', '7', '8', '9']).

c) Eliminación de palabras vacías

Para la eliminación de palabras vacías se utilizó el diccionario en español incorporado en las librerías del NLTK. En la tabla 4.8 podemos observar algunos ejemplos de estas palabras.

un una unas hacemos incluso estado estaba	es soy eres somos atrás porque por qué	por podemos fui fue donde ante antes	unos uno sobre todo también otro algún	fuimos fueron desde ir yo podría cual	estoy esta estamos están como en para	pero quien cuando

Tabla 4.7. En la tabla se presentan ejemplos de palabras vacías, las cuales serán eliminadas de los poemas.

d) Lo poemas se etiquetaron para señalar si correspondían a un autor suicida o no suicida como se muestra en la tabla 4.6.

Ejemplo de texto no suicida	**Ejemplo de texto suicida**
A la señorita julia zarate en la venus de médicis el arte previó cuanto hay en ti, menos la túnica. Irreprochable desnudez imparte al mármol gracia vencedora y única. No te des al acaso. Dios no envía la suprema beldad a cualquier gusto. ¡la manda para ser en la porfía botín al fuerte y galardón al justo! |**nosuicida**	Lied La mañana está de fiesta porque me has besado tú y al contacto de tu boca todo el cielo se hace azul. El arroyo está cantando porque me has mirado tú y en el sol de tu mirada toda el agua se hace azul. El pinar está de luto porque me has dejado tú... y la noche está llorando, noche pálida y azul, noche azul de fin de otoño y de adiós de juventud, noche en que murió la luna, (¡noche en que me has dejado tú!) |**suicida**

Tabla 4.8 La tabla representa un ejemplo de etiquetado de dos poemas del tipo suicida y no suicida; esto se observa al final de cada poema.

e) Los poemas recopilados se organizaron en un corpus donde cada poema fue colocado en una línea, con su respectiva etiqueta a continuación pueden observarse dos textos que representan este proceso.

TEXTO 1:

*¡Está bien! Porque contemplo aún albas radiosas y hay rosas, muchas rosas, muchas rosas en que tiembla el lucero de Belén, y hay rosas, muchas rosas, muchas rosas gracias, ¡está bien! Porque en las tardes, con sutil desmayo, piadosamente besa el sol mi sien, y aun la transfigura con su rayo: gracias, ¡está bien! Porque en las noches una voz me nombra (¡voz de quien yo me sé), y hay un edén escondido en los pliegues de mi sombra: gracias, ¡está bien Porque hasta el mal en mí don es del cielo, pues que, al minarme va, con rudo celo, desmoronando mi prisión también; porque se acerca ya mi primer vuelo: gracias, ¡está bien! |**nosuicida***

TEXTO 2:

*Moderna Yo danzaré en alfombra de verdura, ten pronto el vino en el cristal sonoro, nos beberemos el licor de oro celebrando la noche y su frescura. Yo danzaré como la tierra pura, como la tierra yo seré un tesoro, y en darme pura no hallaré desdoro, Que darse es una forma de la altura. Yo danzaré para que todo olvides y habré de darte la embriaguez qué pides hasta que Venus pase por los cielos. Mas algo acaso te será escondido, que pagana de un siglo empobrecido no dejaré caer todos los velos. |**suicida***

El siguiente ejemplo muestra un poema después del pre-procesamiento:

TEXTO EJEMPLO:

está bien porque contemplo aún albas radiosas y hay rosas muchas rosas muchas rosas que tiembla el lucero de belén y hay rosas muchas rosas muchas rosas gracias está bien las tardes con sutil desmayo piadosamente besa el sol mi sien y aun la transfigura con su rayo gracias está bien las noches voz me nombra voz de quien yo me hay edén escondido en los pliegues de mi sombra gracias está bien porque hasta el mal en mí don es del cielo pues que al minarme va con rudo celo desmoronando mi prisión también se acerca ya mi primer vuelo gracias bien

Al finalizar la etapa de pre-procesamiento el tamaño del vocabulario para este corpus es de 15,962 sin palabras vacías y 16,186 con palabras vacías.

4.3. Representación de los documentos

Una vez pre-procesados los documentos, se procedió a representarlos empleando el modelo vectorial y el esquema de pesado *tf-idf*. Así el poema cuyo texto se muestra pre-procesado previamente, tendría la representación vectorial que se muestra en la figura 4.2. que se simplifica mostrando únicamente los índices con valor diferente de 0, el índice está representado con el color rojo y el valor en color negro.

[54: 0.12068965517241378, 72: 0.06896551724137931,
80: 0.22413793103448276, 84: 0.5689655172413793,
157: 0.034482758620689655, 200: 0.10344827586206896,
205: 3.586206896551724, 210: 0.017241379310344827
216: 0.10344827586206896, 300: 2.086206896551724]

Figura 4.2. Representación tf-idf para un poema pre-procesado clasificado como no suicida.

4.4. Generación del archivo weka

La figura 4.3 nos muestra el diagrama de flujo que representa el proceso del sistema para la generación de los archivos weka.

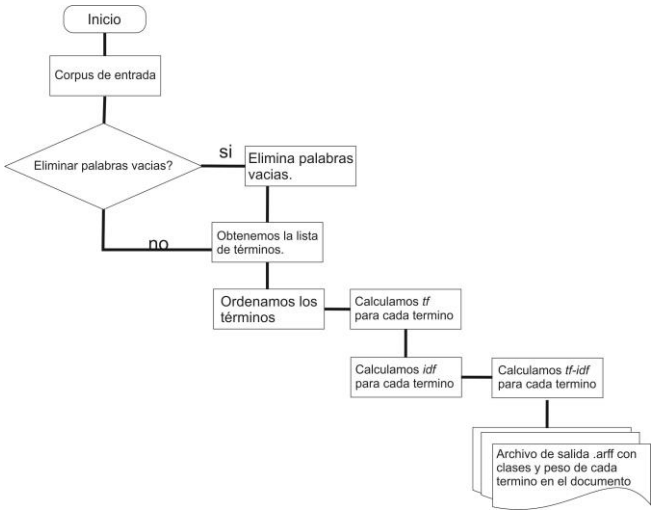

Figura 4.3. Diagrama de flujo que presenta el proceso para la generación del archivo de weka.

En el apéndice 1 pude verse parte de este archivo generado. Cada poema fue representado con datos reales, en el apéndice 2 puede observarse la representación para el poema ¡Esta bien!, del autor Amado Nervo para poemas del tipo no suicida.

Capítulo 5. Experimentos y resultados

En este capítulo se presentan los experimentos realizados con el corpus ya pre-procesado. Se utilizó el software weka 3.8. Empleando los algoritmos Naive Bayes y SVM explicados en el capítulo 2.

5.1. Experimentos

La tabla 5.1 muestra los resultados al utilizar el algoritmo de clasificación SVM y el corpus con palabras vacías, para la creación de los modelos se utilizó validación cruzada a diez pliegues obtenido un porcentaje de precisión de 74.80% para la case no suicida y 73.30% para la clase suicida.

Clase	**Número total de instancias**	**Porcentaje de clasificación**	**Precisión**	**Recuerdo**	**Medida-F**
No suicida	600	74.0196 %	0.7480	0.7400	0.7440
Suicida	600	74.0196 %	0.7330	0.7400	0.7360

Tabla 5.1 Muestra los resultados obtenidos en la evaluación con el algoritmo SVM y corpus con palabras vacías.

La tabla 5.2. Muestra los resultados con el mismo corpus, pero sin palabras vacías y el mismo algoritmo de clasificación, podemos observar que, el porcentaje de clasificación correcta es menor con respecto a la evaluación anterior, obteniendo un porcentaje de precisión del 73.30% para la clase no suicida y 70.90% de precisión para la clase suicida, para este caso suponemos que los resultados decrecen debido a que las palabras vacías capturan aspectos de escritura importantes para la tarea.

Clase	**Número total de instancias**	**Porcentaje de clasificación**	**Precisión**	**Recuerdo**	**Medida-F**
No Suicidas	600	72.0588 %	0.7330	0.7120	0.7220
Suicidas	600	72.0588 %	0.7090	0.7300	0.7190

Tabla 5.2 Resultados de la evaluación obtenidos con el corpus sin palabras vacías y algoritmo de clasificación SVM.

A continuación, se muestra las mismas pruebas, pero evaluadas con el algoritmo de clasificación Naive Bayes en weka.

El algoritmo obtuvo un porcentaje de precisión de 64.2157% con el corpus de palabras vacías, a continuación, se detallan en la tabla 5.3 la precisión, recuerdo y medida-F de esta evaluación.

Clase	**Número total de instancias**	**Porcentaje de clasificación**	**Precisión**	**Recuerdo**	**Medida-F**
No Suicidas	600	64.2157 %	0.6530	0.6350	0.6440
Suicidas	600	64.2157 %	0.6310	0.6500	0.6400

Tabla 5.3 Muestra el resultado de la evaluación con el algoritmo Naive Bayes con el corpus con palabras vacías.

Para el archivo sin palabras vacías el resultado parece haber mejorado dando como resultado un porcentaje de precisión de 65.70% para no suicidas y un 64.60% para la clase suicida.

Clase	**Número total de instancias**	**Porcentaje de clasificación**	**Precisión**	**Recuerdo**	**Medida-F**
No Suicidas	600	66.1765 %	0.6570	0.6630	0.6600
Suicidas	600	66.1765 %	0.6460	0.6400	0.6430

Tabla 5.4 Muestra el resultado de la evaluación con el algoritmo Naive Bayes con el corpus que no contiene palabras vacías.

5.2. Resultados

Para una mejor visualización de resultados se creó una tabla comparativa.

La tabla 5.5 refleja los porcentajes de la evaluación con el corpus de palabras vacías y el de sin palabras vacías con los algoritmos de SVM y Niave Bayes

	Algoritmo SVM		Algoritmo Naive Bayes	
	Porcentaje de precisión		**Porcentaje de precisión**	
Corpus	Con palabras vacías	Sin palabras vacías	Con palabras vacías	Sin palabras vacías
No suicida	74.80%	73.30%	65.30%	65.70%
Suicida	73.30%	70.90%	63.10%	64.60%

Tabla 5.5 Muestra los porcentajes de clasificación correcta para los algoritmos de Naive Bayes y SVM.

La precisión alcanzada por el algoritmo SVM de 74.80% para no suicidas y 73.30% para suicidas, resulta alentadora en comparación al algoritmo de Naive Bayes pues permite comprobar que es posible distinguir tendencias suicidas en los escritos de una persona.

Capítulo 6. Conclusiones y trabajo a futuro

En este trabajo de tesis se construyó un corpus en lenguaje español de 600 documentos con textos de escritores suicidas y no suicidas que fue de utilidad para identificar tendencias suicidas.

La precisión se obtuvo con los procedimientos más sencillos de procesamiento de lenguaje natural: representación de documentos empleando el modelo vectorial y esquema de pesado tf-idf. Estos resultados servirán como referencia para experimentar ahora con otras representaciones tales como: n-gramas, etiquetas sintácticas, vectores de contexto, entre otras. Cabe resaltar que la creación de este corpus de poema en español, es una aportación importante, dada la carencia de recursos lingüísticos que existen en este idioma. La detección de tendencias suicidas a tiempo es de vital importancia para tomar las medidas adecuadas y evitar la pérdida de vidas. Este trabajo constituye un primer paso hacia la detección de estas tendencias suicidas de manera automática.

6.1. Trabajo a futuro

El presente trabajo puede ampliarse para ser utilizado en redes sociales que permita la evaluación de perfiles para identificar tendencias suicidas para generar mecanismos de prevención. Pensando en esta aplicación lo más complicado sería la obtención de ejemplos para entrenar los algoritmos de clasificación.

A partir de esta propuesta, puede pensarse también en el desarrollo de una aplicación que se utilice en el ámbito laborar. Las empresas realizan pruebas psicométricas para evalúan los perfiles de las personas para determinas ciertas cualidades o problemas que estas pueden presentar. La aplicación propuesta podría identificar posibles tendencias suicidas en sus aspirantes y proyectar medidas de prevención.

El corpus generado puede ser utilizado para aplicar diferentes técnicas de procesamiento de lenguaje natural tales como: análisis con n-gramas, frases, signos de puntuación, longitud de palabras y frases, para verificar si el porcentaje de clasificación mejora. Así también como la aplicación de otros algoritmos de clasificación se planea ampliar el corpus creado para poder comprobar si los resultados mejoran.

Bibliografia.

[1] *Estadísticas a Propósito del día Mundial para la Prevención del Suicidio, de INEGI [online]. México: INEGI. 2015 Disponible en: http://www.inegi.org.mx/saladeprensa/aproposito/2015/suicidio0.pdf*

[2] *M. Mulholland, Joanne Quinn. Suicidal Tendencies: The Automatic Classification of Suicidal and Non-Suicidal Lyricists Using NLP [online] International Joint Conference on Natural Language Processing, 2013. Disponible en: http://www.aclweb.org/anthology/I13-1079*

[3] *11º estudio sobre los hábitos de los usuarios de internet en México 2015. 2016, de AMIPCI [online]. AMIPCI. 2015. Disponible en: https://amipci.org.mx/images/AMIPCI_HABITOS_DEL_INTERNAUTA_MEXICANO_2015.pdf*

[4] *John P. Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K. Bretonnel Cohen, John Hurdle, and Christopher Brew. Sentiment Analysis of Suicide Notes: A Shared Task. 2016, de Biomedical Informatics Insights[online]. Junio 2012 Disponible en: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3299408/*

[5] *Garduño Ambriz Rosalía Gómez Hernández Krystiam Yesica Peña Reyes Angélica Julieta. Suicidio en Adolescentes. [online]. Enero 2011. Asociación Mexicana de Tanatología, A.C. Disponible en: http://www.tanatologia-amtac.com/descargas/tesinas/27%20Suicidio%20en%20adolescentes.pdf*

[6] *John P. Pestian, Pawel Matykiewicz, Jacqueline Grupp-Phelan, Sarah Arszman Lavanier, Jennifer Combs, and Robert Kowatch. Using Natural Language Processing to Classify Suicide Notes [online]. 2016, de Cincinnati Children's Hospital Medical Center Cincinnati, OH 45220, USA Disponible en: http://www.aclweb.org/anthology/W08-0616*

[7] *Azadeh Nikfarjam, Ehsan Emadzadeh, and Graciela Gonzalez. A Hybrid System for Emotion Extraction from Suicide Notes [online]. 2016. Biomed Inform Insights Disponible en: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3409484/*

[8] *Hui Yang, Alistair Willis, Anne de Roeck, and Bashar Nuseibeh. A Hybrid Model for Automatic Emotion Recognition in Suicide Notes [online]. 2016. Biomed Inform Insights Disponible en: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3409477/*

[9] *Irena Spasić, Pete Burnap, Mark Greenwood, and Michael Arribas-Ayllon. A Naïve Bayes Approach to Classifying Topics in Suicide Notes [online]. 2012. Biomed Inform Insights Disponible en: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3409485/*

[10] *James A. McCart, Dezon K. Finch, Jay Jarman, Edward Hickling, Jason D. Lind, Matthew R. Richardson, Donald J. Berndt, and Stephen L. Luther. Using Ensemble*

Models to Classify the Sentiment Expressed in Suicide Notes [online]. 2012. Biomed Inform Insights, Disponible en: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3409473/

[11] *Introducción a la Recuperación de Información [online]. Disponible en http://www.grupolys.org/docencia/ln/biblioteca/ir.pdf*

[12] *INEGI. ESTADÍSTICAS A PROPÓSITO DEL DÍA MUNDIAL PARA LA PREVENCIÓN DEL SUICIDIO [online]. 2016, INEGI Disponible en: http://www.inegi.org.mx/saladeprensa/aproposito/2016/suicidio2016_0.pd*

[13] *Diego García Morate. MANUAL DE WEKA. 20 [online]. 2000. Creative Commons Reconocimiento-NoComercial-SinObraDerivada 2.0 Disponible en: http://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/InteligenciaDeNegocio/weka.pdf*

[14] *P. F. Machine learning. Cambridge University Press 2012*

[15] *Blázquez M. Técnicas avanzadas de recuperación de información [online]. 2012. Investigación en Documentación Disponible en: http://ccdoc-tecnicasrecuperacioninformacion.blogspot.mx/2012/11/frecuencias-y-pesos-de-los-terminos-de.html*

[16] *C. Morales, "Clasificación Automática de Textos considerando el estilo de redacción", M.S.tesis, INAOE, Puebla, 2007.*

[17] *Organization mundial de la salud [online]: OMS Disponible en:http://www.who.int/mediacentre/factsheets/fs398/es/*

[18] *Gabriel H., Fernando R. Introducción a la Recuperación de Información [online]. Universidad Nacional de Luján. Disponible en: http://eprints.rclis.org/12243/1/Introduccion-RI-v9f.pdf*

[19] *Aprendizaje Supervisado y no Supervisado 2013 [online] Disponible en: http://redesneuronares.blogspot.mx/*

Apéndice 1

Estructura y formato de un archivo .arff de weka

Formatos weka

Weka es una extensa colección de algoritmos de máquinas de conocimiento desarrollados por la universidad de Waikato (Nueva Zelanda) implementados en java; Útiles para ser aplicados sobre datos mediante interfaces que ofrece. Weka contiene las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, clustering, asociación y visualización. [13]

Ficheros .arff

Nativamente Weka trabaja con un formato denominado arff, acrónimo de *Attribute-Relation File Formal.* Este formato está compuesto por una estructura claramente diferenciada en tres partes.

1. Cabecera. Se define el nombre de la relación
   Formato
   @relation <nombre-de-la-relación>
   Donde <nombre-de-la-relación> es de tipo String. SI dicho nombre contiene algún espacio será necesario expresarlo entrecomillado.

2. Declaraciones de atributos. En esta sección se declaran los atributos que compondrán nuestro archivo junto a su tipo.
   Formato
   @attribute <nombre-del-atributo> <tipo>
   Donde <nombre-del-atributo> es de tipo String teniendo las mismas restricciones que el caso anterior.
   Tipos:
   a) Numeric Números reales.
   b) Integer Números enteros
   c) Date Fechas, para ello este tipo debe ir precedido de una etiqueta de formato entrecomillada. La etiqueta de formato está compuesta de caracteres separadores (guiones y/o espacios) y unidades de tiempo:
      - dd Día
      - MM Mes

- yyyy Año
- HH Horas
- Mm Minutos
- Ss Segundos

d) String Expresa cadenas de texto

e) Enumerado El identificador de este tipo consiste en expresar entre llaves y separados por comas los posibles valores (caracteres o cadenas de caracteres) que puede tomar el atributo. Por ejemplo, si tenemos el atributo que indica el tiempo podría definirse;
@attribute tiempo {soleado, lluvioso, nublado}

3. Sección de Datos. Declaramos los datos que componen la relación separando entre comas los atributos y con saltos de línea las relaciones
Ejemplo
@data

4,3.2

Apéndice 2

A continuación, se presenta parte del contenido de nuestros archivos .arff de weka para el corpus con palabras vacías y el corpus sin palabras vacías respectivamente, en ellos se puede identificar el formato explicado en el apéndice 1.

Cabecera, que define que es un archivo con palabras vacías

@RELATION CuerpoCW

Atributos, que representan a cada una de nuestras palabras en el documento

@ATTRIBUTE a REAL

@ATTRIBUTE abad REAL

@ATTRIBUTE abadía REAL

@ATTRIBUTE abajo REAL

@ATTRIBUTE abandona REAL

@ATTRIBUTE abandonaba REAL

@ATTRIBUTE abandonada REAL

@ATTRIBUTE abandonadas REAL

@ATTRIBUTE abandonado REAL

@ATTRIBUTE abandonados REAL

@ATTRIBUTE abandonando REAL

@ATTRIBUTE abandonar REAL

@ATTRIBUTE abandonarme REAL

@ATTRIBUTE abandonarse REAL

@ATTRIBUTE abandonará REAL

@ATTRIBUTE abandonas REAL

@ATTRIBUTE abandones REAL

@ATTRIBUTE abandono REAL

@ATTRIBUTE abandonos REAL

@ATTRIBUTE abandoné REAL

@ATTRIBUTE abandonó REAL

@ATTRIBUTE abanica REAL

@ATTRIBUTE abanico REAL

Seccion de datos

@DATA

0.0, 0.06140350877192982, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2.9298245614035086, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.03508771929824561,

Para nuestro documento sin palabras vacías

@RELATION CuerpoSW

@ATTRIBUTE abad REAL

@ATTRIBUTE abadía REAL

@ATTRIBUTE abajo REAL

@ATTRIBUTE abandona REAL

@ATTRIBUTE abandonaba REAL

@ATTRIBUTE abandonada REAL

@ATTRIBUTE abandonadas REAL

@ATTRIBUTE abandonado REAL

@ATTRIBUTE abandonados REAL

@ATTRIBUTE abandonando REAL

@ATTRIBUTE abandonar REAL

@ATTRIBUTE abandonarme REAL

@ATTRIBUTE abandonarse REAL

@ATTRIBUTE class {nosuicida,suicida}

@DATA

0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.12068965517241378, 0.0, 0.0, 0.22413793103448276, 0.0, 0.5689655172413793, 0.034482758620689655.