UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Tarea 1 Preprocesamiento de Datos

Abril Grisel Guevara Cedillo
Facultad de Ciencias Físico Matemáticas

Introducción

Para este trabajo se solicitó realizar el preprocesamiento de información sobre un tema que pudiera relacionarse con mi proyecto de tesina por lo que elegí el tema de Suicidios debido a que, aunque no tengo definido aún el tema de mi tesina, me interesa que este relacionado a un tema social. La base de datos que utilizaré se encuentra en Kaggle es una colección de publicaciones de los subreddits "SuicideWatch"y "Depresión"de la plataforma Reddit, las publicaciones se recopilan utilizando la API pushshift. Todas las publicaciones que se incluyeron se hicieron en "SuicideWatch"desde el 16 de diciembre de 2008 (creación) hasta el 2 de enero de 2021, se recopilaron, mientras que las publicaciones de "Depresión"se recopilaron desde el 1 de enero de 2009 hasta el 2 de enero de 2021. En la base también se incluye una columna con la información de ideación suicida, la cual es la tendencia a pensar de manera repetida en la posibilidad de terminar con la propia vida. Puede presentarse en diversos grados de intensidad o severidad, desde la ideación suicida pasiva, en la que la persona cobra consciencia de que no quiere seguir viviendo, hasta la ideación suicida, en la que piensa acerca de distintas alternativas para terminar con su vida.

Objetivo

El objetivo de esta trabajo es realizar el preprocesamiento de la base de datos mencionada para realizar un comparativo y determinar si existe alguna diferencia en las palabras más frecuentes en los comentarios de personas con ideación suicida y las que no lo presentan.

Metodología

Para realizar el preprocesamiento realicé 5 pasos:

- Extracción de los comentarios
- Tokenización
- Conversión de palabras a minúsculas
- Separar Stopwords de las No Stopwords
- Lematización

A continuación se explica cada punto, después de la extracción del texto de los comentarios utilicé la librería de lenguaje natural NLTK para realizar la tokenización la cual consiste en separar el texto palabra por palabra. Al obtener estas palabras, las convertí en minúsculas para estandarizar las palabras. Después obtuve las stopwords, las cuales son palabras que tienen una frecuencia alta debido a que son palabras conectoras. A las palabras que quedan al quitar las stopwords se les aplique la lematización, la cual consiste en agrupar las diferentes formas flexionadas de una palabra para que puedan analizarse como un solo elemento. Al final realicé una gráfica de la distribución de las frecuencias de las palabras y una nube de palabras para los comentarios que presentaban ideación suicida y para las que no lo presentaban para de esta manera realizar un comparativo.

Resultados

A continuación se muestran las gráficas obtenidas

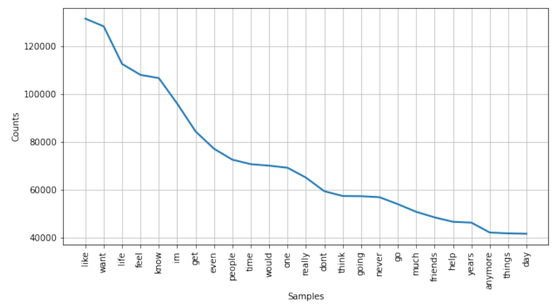

Figura 1:Distribución de frecuencias Ideación de Suicidio

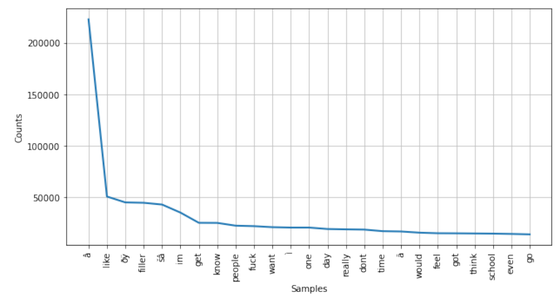

Figura 2:Distribución de frecuencias No ideación de Suicidio

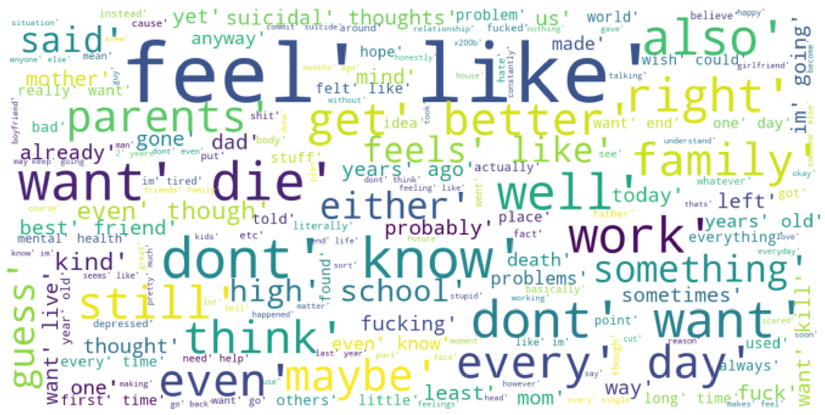

Figura 3:Nube de palabras Suicidios

Figura 4:Nube de palabras No Suicidios

Conclusión

Los comentarios de las personas que presentan ideación de suicidio muestra palabras como: die, suicidal, problem, feel, mental health, death, las cuales no se presentan en los comentarios de las personas que no lo presentan. Esto podría respaldar que la afirmación de que las personas que se quitan la vida no dan antes ningún tipo de señal es un mito.

Referencias

- https://github.com/AbrilGuevara/Procesamiento-de-Datos
- https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch?resource=download
- https://www.avancepsicologos.com/ideacion-suicida/