Universidad Autónoma de Nuevo León
Facultad de Ciencias Físico Matemáticas
Maestría en Ciencia de Datos

Tarea 2

-ANÁLISIS DE SENTIMIENTO-
REDDIT SUICIDIOS

Procesamiento y Clasificación de Datos
MET. Mayra Cristina Berrones Reyes

Alumna:
Abril Grisel Guevara Cedillo

Matrícula:
1419239

San Nicolás de los Garza a 28 de mayo del 2022

0.1. Objetivo

Realizar un análisis de sentimiento, haciendo comparación entre las diferentes librerías que se mencionaron en clase 2. Discutir y analizar los resultados para concluir cual es la mejor librería de las utilizadas.

0.2. Introducción

Continuaré con la misma base de datos que utilicé para la tarea 1, la cual se encuentra en Kaggle es una colección de publicaciones de los subreddits "SuicideWatchz "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.

0.3. Procesamiento de los datos

Antes de aplicar el análisis de sentimientos, realicé el procesamiento de los datos el cual cambié un poco versus el utilizado en la tarea 1, me basé en el procesamiento visto en la clase, a continuación detallaré los pasos aplicados para el procesamiento de la base.

- Crear un data frame con los post de Reddit en cada uno de los renglones.
- Limpiar la base

  Quitar caracteres especiales y números de la base mediante una función definida como clean, estos datos limpios se agregan al final del data frame en una columna llamada Cleaned Posts.
- Tokenización de los posts y stop words

  Usando la librería NLTK se separaron los posts palabra por palabra, se quitaron los stopwords y se asignó a cada una de las palabras que tipo de palabra son en base a un diccionario:

  - 'J' a los adjetivos
  - 'V' a los verbos
  - 'N' a los sustantivos
  - 'R' a los adverbios

  Al final del data set se agrega una columna llamada POS tagged con los resultados de este paso

- Lematización

  Con la misma librería NLTK se aplica la lematización de los posts ya tokenizados para obtener la raíz léxica de las palabras. Al final se crea un data frame de los post en una columna con los post lematizados en la otra columna.

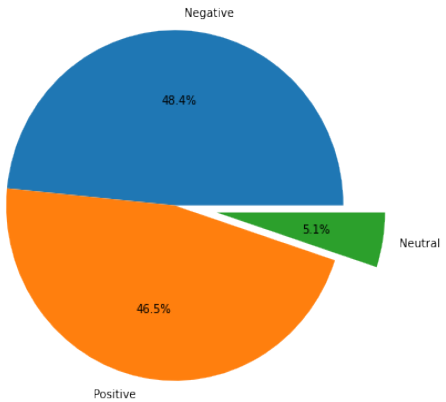
0.4. Análisis de sentimiento

El análisis de sentimientos, también conocido como minería de opiniones, es un término muy difundido, pero a menudo mal entendido. En esencia, es el proceso de determinar el tono emocional detrás de una serie de palabras, que se utiliza para comprender las actitudes, opiniones y emociones expresadas en una mención en línea. Para los enfoques basados en el léxico, un sentimiento se define por su orientación semántica y la intensidad de cada palabra en la oración.

Se aplicaron las 3 librerías mencionadas en la clase:

- Análisis de sentimiento usando TextBlob

  Se crearon 2 funciones, una para calcular la polaridad de cada post y otra para etiquetar el post con la tipo de sentimiento. La propiedad de sentimiento de TextBlob devuelve un objeto de sentimiento. La polaridad indica sentimiento con un valor de -1.0 (negativo) a 1.0 (positivo) con 0.0 siendo neutral. La subjetividad es un valor de 0.0 (objetivo) a 1.0 (subjetivo), pero en este caso no utilicé la subjetividad, la segunda función es para etiquetar cada uno de los post en un sentimiento siendo menores a cero los negativos, mayores a cero los positivos y cero los neutrales. Al final se contabiliza cada uno de los sentimientos para obtener la frecuencia en una gráfica de pay.

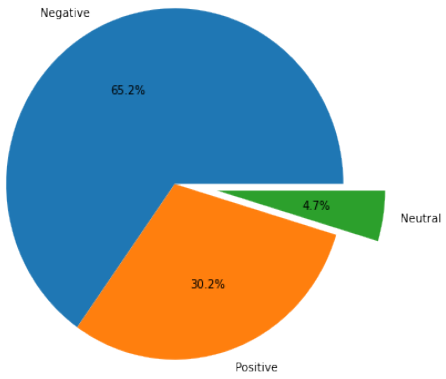


Gráfica 1: Análisis de sentimiento usando TextBlob

Los resultados muestran un porcentaje más alto en el sentimiento negativo 48.4 %, seguido por un 46.5 % en el sentimiento positivo y un 5.1 % neutral.

- Análisis de sentimiento usando VADER

  Se creo una función para calcular el sentimiento basado en esta librería, Vader utiliza una lista de características léxicas por ejemplo una palabra que se etiquetan como positivas o negativas según su orientación semántica para calcular el sentimiento de texto. La función devuelve la probabilidad de que una oración de entrada dada, sea positiva, negativa o neutral. La suma de las 3 probabilidades da como resultado un 1. Devuelve la probabilidad compuesta çompound"por cada uno de los post. Se define un umbral para saber de cual valor a cual valor de compound se etiquetara cada uno de los sentimientos. En el ejemplo visto en clase los umbrales fueron de -.5 a .5 para neutral, como el rango es amplio había un porcentaje importante de posts neutrales por lo que decidí cambiar el umbral de clasificación de los sentimientos de -.1 a 1 haciendo el rango muy pequeño casi en el cero, por lo tanto todos los posts con compound menor a -.1 son etiquetados como negativos y los compound mayores a .1 son etiquetados como positivos. Al final se contabiliza cada uno de los sentimientos para obtener la frecuencia en una gráfica de pay.



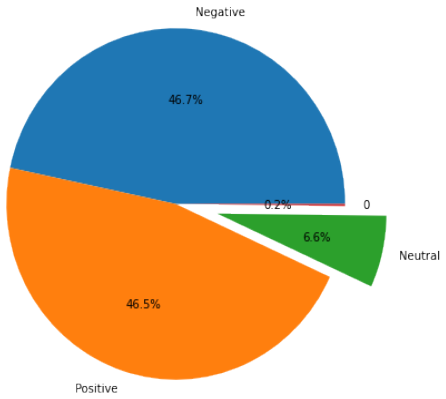
Gráfica 2: Análisis de sentimiento usando VADER

Los resultados muestran un porcentaje más alto en el sentimiento negativo 65.2 %, seguido por un 30.2 % en el sentimiento positivo y un 4.7 % neutral.

- Análisis de sentimiento usando SentiWordNet

  La librería utiliza la base de datos WordNet. Es importante obtener el lema de cada palabra, se obtienen puntuaciones objetivas, positivas y negativas para todos los sintetizadores posibles y se etiqueta cada uno de los

posts. Si la puntuación positiva es mayor que la negativa, el sentimiento es positivo, si la puntuación positiva es menor que la puntuación negativa, el sentimiento es negativo y si la puntuación positiva y negativa son iguales, el sentimiento es neutral. Al final se contabiliza cada uno de los sentimientos para obtener la frecuencia en una gráfica de pay.

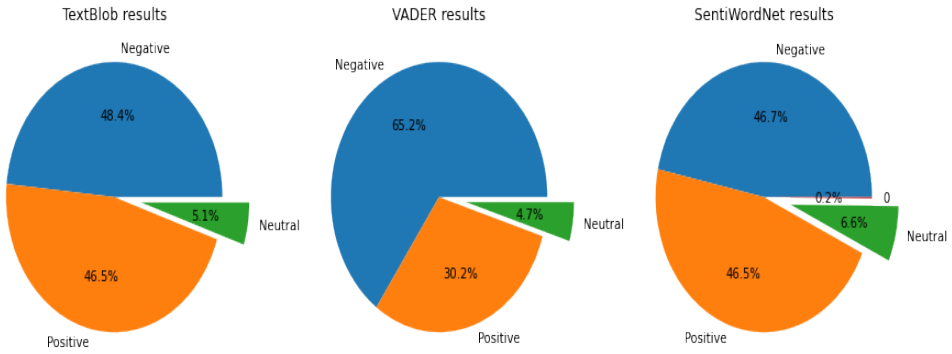


Gráfica 3: Análisis de sentimiento usando SentiWordNet

Los resultados muestran un porcentaje apenas visiblemente más alto en el sentimiento negativo 46.7 %, seguido por un 46.5 % en el sentimiento positivo y un 6.6 % neutral.

0.5. Resultados y Conclusiones

En la gráfica comparativa de todos los pays resultado de las 3 librerías aplicadas, puedo observar que las librerías TextBlob y SentiWordNet son muy parecidos entre ellos no se aprecia un sesgo hacia alguno de los 2 sentimientos. La librería VADER muestra mejores resultados debido a que si hay una categoría con mayoría de proporción, con esta librería el sentimiento que más predomina en los posts es el negativo como era de esperarse debido al contexto de la base de datos. Creo que lo que la hace tener una ventaja es que los analistas podemos definir desde cual y hasta cual valor van a ser los umbrales de decisión del compound obtenido por la librería. Debido a que en este tipo de análisis se busca encontrar el sentimiento al que apunta el post es importante que pocos posts sean clasificados como neutrales, por lo que funcionó haber acotado el rango de decisión para las etiquetas de los sentimientos.



Gráfica 4: Comparativo de resultados entre librerías

0.6. Referencias

Análisis de sentimiento con TextBlob y Python
https://www.linuxteaching.com/article/sentiment_analysis_with_textblob_and_python#what_is_difference_between_nltk_and_textblob
Textblob vs Vader para análisis de sentimientos en Python.
https://datapeaker.com/big-data/guia-para-el-procesamiento-del-lenguaje-natural-en-pytho
Análisis de sentimientos basado en reglas en Python para científicos de datos
https://datapeaker.com/big-data/analisis-de-sentimientos-basado-en-reglas-en-python-para
Github https://github.com/AbrilGuevara/Procesamiento-de-Datos/tree/main/Tarea%202
Base de datos https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch?resource=downloadhttps://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch?resource=download