Detección de intentos de suicidio

El culto de Dijkstra:

- Gabriel Beltrán
- Nicolás Fajardo
- Daniel Zambrano

Comprensión del Negocio

Entendimiento del negocio

- Reddit es una red social que permite a los usuarios subscribirse a varias comunidades conocidas como ‘subreddit’.
- Cada uno de estos ‘subreddits’ están centrados a un tema, que puede ser bastante general (como r/movies que está centrado en películas) o específico (como r/DunderMifflin que está centrado en la popular serie ‘The Office (US)’).

Comprensión del Negocio

Oportunidad/problema del negocio

- Las comunidades r/Depression y r/SuicideWatch están centradas en que sus miembros se dan apoyo mutuo en salud mental, para aquellos que tengan pensamientos depresivos y/o suicidas.
- Entre los usuarios de esas comunidades, hay quienes han intentado quitarse la vida, incluso algunos que lo hayan logrado, aunque es difícil de saber.
- Una oportunidad para el negocio puede ser la de buscar identificar usuarios que pueden ser propensos a intentar quitarse la vida, para así tratar de brindarles más asistencia o insistirles que busquen ayuda profesional.

Comprensión del Negocio

Objetivos y criterios de éxito

- Como objetivo se tiene el poder identificar qué usuarios son propensos a intentar quitarse la vida.
- Para saber si se tiene éxito no basta con realizar una predicción, sino que toca buscar una predicción correcta.
- No es necesario llegar a un 100% de precisión, sino que solo basta con llegar a un 0% de falsos negativos, pues de ese modo se puede asegurar que por lo menos todos los que sean propensos a intentar quitarse la vida sean identificados y que reciban ayuda, así se identifiquen erróneamente usuarios que en realidad no sean propensos.

Comprensión del Negocio

Enfoque analítico

- Creación de un modelo de clasificación de usuarios de la red social de acuerdo a si son propensos a intentar quitarse la vida mediante un proceso de análisis de sentimientos usando algoritmos de clasificación.
- Al desarrollar un modelo que pueda clasificar a los usuarios entre aquellos que sean propensos a quitarse la vida y los que no, es posible dar cumplimiento al objetivo de negocio de precisamente saber a qué usuarios es necesario brindar una mayor asistencia en salud mental.

Comprensión del Negocio

Organizaciones que se benefician

- Reddit: dueños y administradores/moderadores de los subreddits r/SuicideWatch y r/depression. Se benefician al poder identificar usuarios que sean propensos a intentar quitarse la vida.
- Fundaciones y ONG de prevención del suicidio: estas pueden brindarle una primera atención a los usuarios identificados y, de ser el caso, remitirlos a una ayuda profesional, para así cumplir con su misión organizacional.
- Centros de atención psicológica: pueden recibir pacientes y brindarles un apoyo profesional.

Comprensión del Negocio

Tecnologías y algoritmos a usar

- Algoritmos de clasificación:
  - Random Forest.
  - RoBERTa.
  - Logistic Regression.
- Herramientas de tokenización y vectorización:
  - RedditTokenizer.
  - TF-IDF.

Entendimiento y Preparación de los datos

Perfilamiento de los datos

- Los datos consisten en una tabla con 3 columnas: un id de la fila, el comentario de un usuario y un valor binario que indica si ese usuario se ha intentado quitar la vida.
- Solo se conocen los datos de quienes han intentado quitarse la vida, y lo reportaron de algún modo. Es posible que la tabla misma presente falsos negativos.
- Hay un ruido importante en caso de que un usuario que haya intentado suicidarse tenga más de un comentario en la base de datos.

Entendimiento y Preparación de los datos

Análisis de la calidad de los datos

- En total, hay 17 filas con comentarios nulos, y no hay filas con la clasificación nula.
- En cuanto a la integridad de los datos, la columna de comentarios, una alta cantidad de estos presenta diferentes caracteres que pueden no ser ideales para el análisis, como caracteres especiales, de puntuación, numéricos e incluso emojis.
- En cuanto a la columna de la clasificación

Entendimiento y Preparación de los datos

Tratamiento de los datos

- Se realiza una limpieza inicial mediante Tableau:
  - Eliminación de registros nulos.
  - Remover caracteres especiales, numéricos, de puntuación y emojis.
- Se realiza un proceso de tokenización y vectorización:
  - Se usan TF-IDF y BoW, aunque este último se descarta por la gran cantidad de memoria que usa.
  - Se usa un tokenizer conocido como RedditTokenizer.

Modelado y evaluación

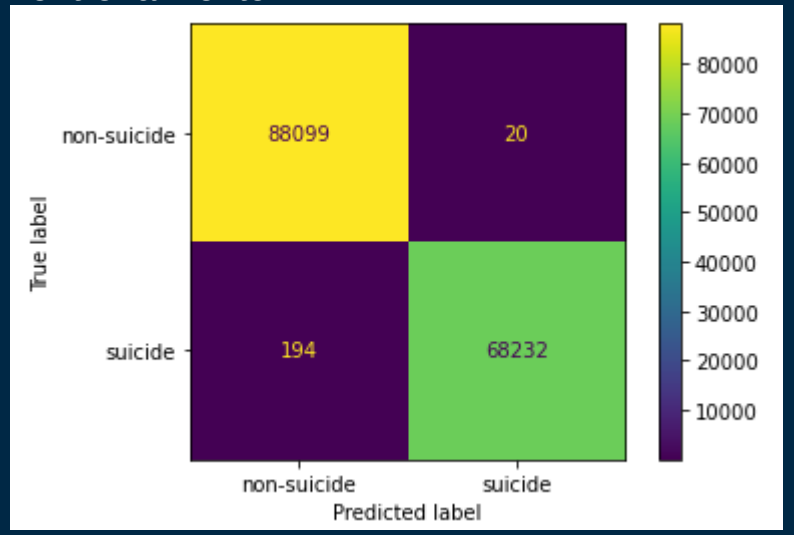
Modelos usados

- Random Forest
- RoBERTa
- Logistic Regression

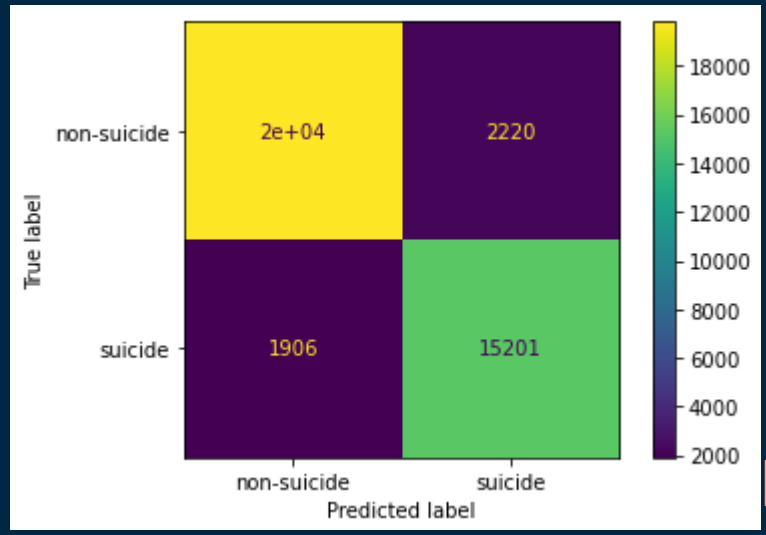
Random Forest

- Inicialmente se seleccionan como hiperparámetros un random_state = 3 y n_estimators = 500.
- Para el conjunto de entrenamiento se obtiene una precisión, exhaustividad y f1 de: 0.9997, 0.9972 y 0.9984, respectivamente.
- Para el conjunto de entrenamiento se obtiene una precisión, exhaustividad y f1 de: 0.8726, 0.8886 y 0.8805, respectivamente.

Matriz de confusión para el conjunto de entrenamiento

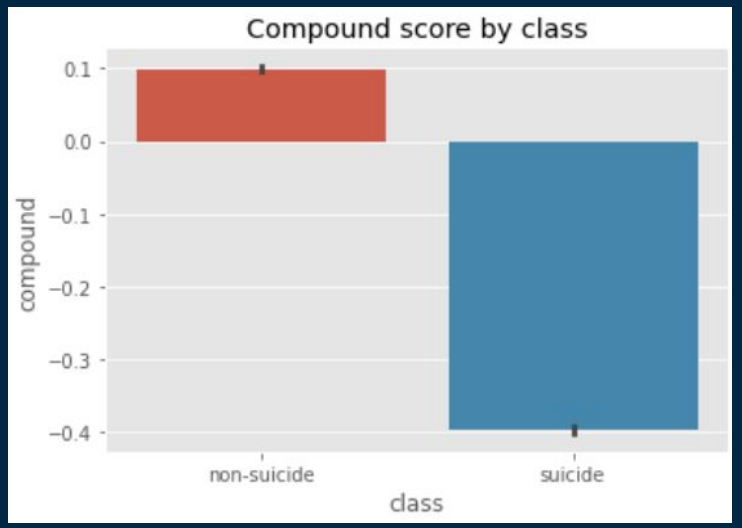


Matriz de confusión para el conjunto de prueba



RoBERTa

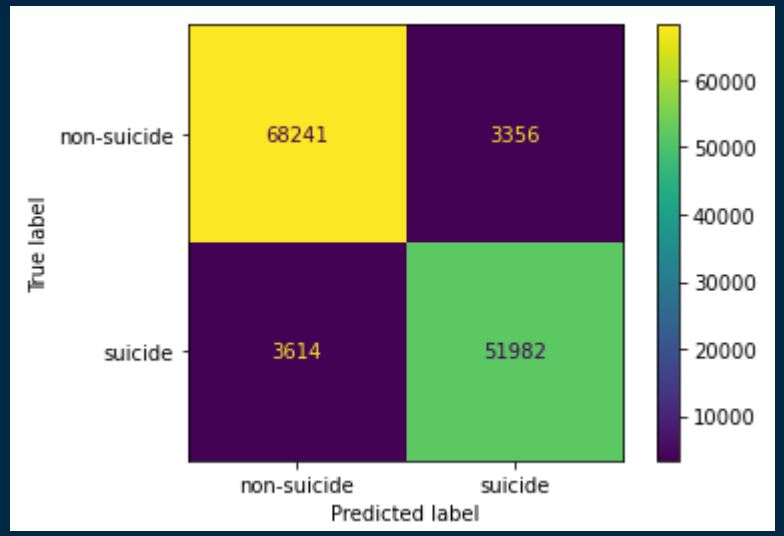
- Se parte de usar VADER, una herramienta de análisis de sentimientos entrenada previamente con sentimientos de redes sociales.
- Se usa el modelo RoBERTa, uno pre-entrenado con tareas de MLM. Este modelo considera la parte inferencial de los textos al estar entranada en base a transformadores.
- Presenta limitaciones debido a la extensión de los textos.



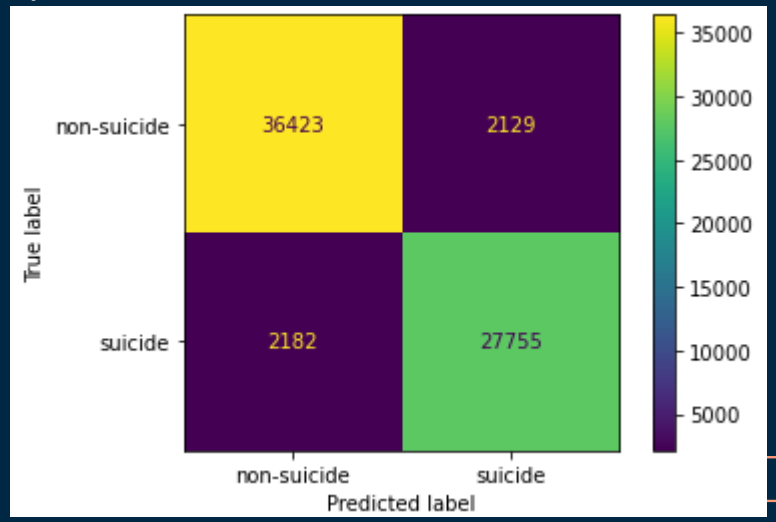
Logistic Regression

- Se selecciona como hiperparámetro un class_weight = ‘balanced’.
- Para el conjunto de entrenamiento se obtiene una precisión, exhaustividad y f1 de: 0.9394, 0.9350, 0.9372, respectivamente.
- Para el conjunto de entrenamiento se obtiene una precisión, exhaustividad y f1 de: 0.9288, 0.9271, 0.9279, respectivamente.

Matriz de confusión para el conjunto de entrenamiento



Matriz de confusión para el conjunto de prueba



Resultados

- RoBERTa es un modelo considerablemente robusto, pero está optimizado para textos cortos, por lo que puede no ser muy apropiado para esta tarea.
- Los otros dos modelos, Random Forest y Logistic Regression, presentaron resultados prometedores.
- El modelo de Random Forest presentó para el conjunto de entrenamiento un valor de f1 bastante alto (99.8%), y pocos falsos negativos (194 entre los 150000 datos del conjunto de pruebas). Para el conjunto de prueba, se obtuvo un f1 de 88.1%, que es bastante alto, y un 5% de falsos negativos.
- El modelo de Logistic Regression presentó para el conjunto de entrenamiento un valor de f1 de 93.7% y una cantidad considerable de falsos negativos, 3614 entre 130000 datos que fueron usados en este conjunto. En el caso del conjunto de prueba, se cuenta con un valor de f1 un poco menor, 92.8%, pero tuvo unos 2100 falsos negativos, un 3% de falsos negativos.