**Mejora Hito 1**

A partir de los resultados obtenidos en el Hito 1, se plantearon en su momento las siguientes preguntas/hipótesis. Algunas de ellas ya han sido respondidas, algunas serán evaluadas en el hito 3 y otras serán eliminadas del proyecto según se indica a continuación.

1. ¿El sentimiento general sobre el COVID-19 varía por la localidad registrada?

A partir del análisis exploratorio de datos se respondió esta pregunta, en el presente hito se procedió a estandarizar los valores de localidad que tenia la base de datos inicial, para de esta manera agrupar los datos de forma mas ordenada, para que los resultados obtenidos permitan una mejor visualización y comprensión de los datos. No es necesario evaluar esta pregunta en el hito 3.

1. Dada las características del COVID-19 y sus consecuencias, más del 50% de los tweets están asociados a un sentimiento negativo o extremadamente negativo.

Esta pregunta fue respondida para la base de datos inicial en el Hito 1 sin embargo se incorporará nueva información que necesita ser evaluada (set de evaluación). Se comparan los análisis de sentimiento obtenidos para los dos grupos de información.

Si hubo alguna variación de sentimiento en estos dos instantes de tiempo.

1. ¿Los sentimientos mayormente expresados en los Tweets tienen relación con el contexto social del lugar desde donde son publicados?

Debido a la orientación que se le dará finalmente al proyecto esta pregunta se encuentra fuera de los alcances, por lo tanto, se eliminará.

1. ¿Se puede asociar un sentimiento a una palabra dependiendo de las otras palabras que forman el tweet analizado?

No se evaluará debido a la complejidad de las herramientas necesarias para abordar el problema, estas no se encuentran en el contenido del curso.

1. ¿Existe una variacion de comentarios positivos al avanzar los días? ¿Si es así, a qué se debe?

No es posible establecer una razón clara a la cual atribuir la variación de sentimiento en los tweets.

1. Cambiar a hipótesis, cambia la percepción de las personas a partir de marzo

¿Se puede generar un clasificador que se pueda generalizar en base a los datos que tenemos?

La pregunta será reemplazada por la siguiente hipótesis:

Los sentimientos de una nueva base de datos pueden ser definidos a partir de un clasificador entrenado con una base de datos previa.

La metodologia a usar para comprobar esta hipótesis se describe en el capitulo propuesta experimental.

1. ¿El sentimiento de los comentarios se ha visto modificado con respecto a tiempos anteriores al Covid?

No es posible encontrar una base de datos, previa a la pandemia, que sirva como comparación al problema que se aborda en el presente análisis; por lo tanto, esta pregunta no se responderá.

**Propuesta experimental**

Con el fin de responder las preguntas e hipótesis que quedan por evaluar, se propone seguir la metodología.

* En el data set inicial existen ubicaciones desde donde se escribieron los Tweets descritas por ciudad, para el proyecto las ubicaciones deben ser dadas por país, así que se estandarizó este atributo de manera de poder interpretar la información. También se eliminaron hashtags, pues no son considerados en el análisis de sentimiento.
* Se extraerá características del texto de los tweets representándolo de forma vectorial para relacionarlo con el sentimiento que se le ha asignado de forma manual.
* Se dividirá la base de datos en dos grupo set de entrenamiento y set de evaluación, con un porcentaje de 80% y 20% respectivamente .
* Para definir el modelo que arroja una mejor clasificación se evaluará la base de datos usando Naive Bayes, Arboles de desicion, K\_Mean y SVM. Se asignaran hiperparametros de evaluación a cada uno de los modelos a usar.
* Los resultados obtenidos al aplicar los diferentes modelos se comparar usando métricas tradicionales F1, precision y recall, además aplicando validación cruzada sobre los grupos cross validation sobre lo grupos de entrenamiento y testeo respectivamente.
* Después de elegir el modelo que mejor clasifica los datos, se usará para asignar sentimiento a una base de datos nueva que será limpiada y estandarizada según los atributos de la base de dato inicial.
* Se agrupará la asignación de sentimiento de la base de datos nueva según ubicación y se comparará con la base de datos inicial, para así identificar si existe alguna tendencia.