# Motivacion:

Twitter es una de las redes sociales más utilizadas para comentar, compartir o debatir temas de actualidad y tendencias. El primer trimestre de 2020 tuvo un aumento de un 23% en comentarios diarios con respecto al mismo periodo de tiempo del año 2019 [1], siendo una de sus principales causas la pandemia del COVID-19.

Considerando este contexto, es que analizar datos de Twitter es interesante, pues además de que es posible obtener información en tiempo real de sucesos importantes, como el caso de la pandemia, se pueden aplicar métodos de minería de datos para manipular grandes volúmenes de información, analizándola y obteniendo en muchas ocasiones conclusiones no triviales acerca de esta.

En el presente estudio se explorarán entonces Tweets relacionados con coronavirus en un intervalo de tiempo acotado, desde la perspectiva de los sentimientos y relacionándolos con el contexto país desde donde se emiten; buscando establecer si este ultimo influencia la percepción de las personas acerca de la pandemia.

Es importante aclarar que el procesamiento del lenguaje natural como la herramienta a utilizar debido a la indiscutible importancia que supone identificar la subjetividad y el grado de valoración de miles de personas en cada instante para empresas, organizaciones, gobiernos y la población misma.

Dicho lo anterior, los puntos claves a analizar en el desarrollo del proyecto son:

- Identificar cómo se relaciona el sentimiento identificado con el contexto país (según mayor o menor presencia del sentimiento).

- Identificar palabras que son clave a la hora de categorizar el sentimiento.

- Establecer algoritmos para predecir sentimientos de forma sistematizada.

- Entrenar modelos de clasificaciòn en base a tweets usando un dataset de entrenamiento y un dataset de evaluacion.

# Descripción de base de datos

La base de datos, extraída desde la plataforma \*Kaggle\*[2], esta compuesta en principio por 44.955 tweets relacionados con el tema COVID 19 y que fueron publicados desde el 2 de marzo al 14 de abril de 2020. Además de encontrarse el texto publicado se encuentran en la base de datos los atributos de fecha exacta de publicación, ubicación desde la cual se realizla publicación, un identificador para el usuario y la asignación de sentimiento para cada tweet. La asignación de sentimiento a cada Tweet fue realizada de forma manual por el propietario de la base de datos. Esta variable de sentimiento sería un punto de comparación para un posible modelo de clasifiación de los sentimientos de los tweets.

Dicho lo anterior, los puntos claves a analizar en el desarrollo del proyecto son:

- Identificar cómo se relaciona el sentimiento identificado con el contexto país (según mayor o menor presencia del sentimiento).

- Categorizar países según mayor o menor presencia de sentimiento y relacionarlos con algun índice de felicidad publicado en el último año.

- Identificar palabras que son clave a la hora de categorizar el sentimiento.

- Establecer algoritmos para predecir sentimientos de forma sistematizada.

- Entrenar modelos de clasificaciòn en base a tweets usando un dataset de entrenamiento y un dataset de evaluacion.

# Descripción de base de datos

La base de datos, extraída desde la plataforma \*Kaggle\*[2], esta compuesta en principio por 44.955 tweets relacionados con el tema COVID 19 y que fueron publicados desde el 2 de marzo al 14 de abril de 2020. Además de encontrarse el texto publicado se encuentran en la base de datos los atributos de fecha exacta de publicación, ubicación desde la cual se realizla publicación, un identificador para el usuario y la asignación de sentimiento para cada tweet. La asignación de sentimiento a cada Tweet fue realizada de forma manual por el propietario de la base de datos. Esta variable de sentimiento sería un punto de comparación para un posible modelo de clasifiación de los sentimientos de los tweets.