

# Hito 3: Análisis de sentimiento en Tweets relacionados con Covid-19

**Nicolás Herrera - Yesenia Marulanda - Franco Migliorelli  
Samuel Sánchez - Sebastián Urbina  
Grupo 11**

06 de Enero de 2021  
CC5206 - Introducción a la Minería de Datos

# AGENDA

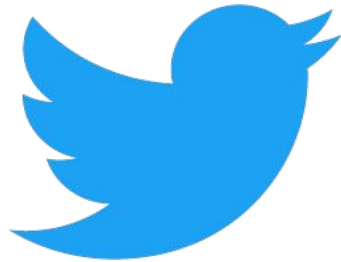
1. Introduccion
2. Preguntas/hipotesis
3. Metodologia.
4. Experimentos y resultados.
  - 4.1. Análisis exploratorio de los datos.
  - 4.2. Preprocesamiento.
  - 4.3. Vectorización de comentarios.
  - 4.4. Predicción sobre un nuevo dataset.
5. Analisis futuros.



# 1. Introducción

Motivación

# 1. INTRODUCCIÓN



**+23%**

tweets diarios en primer  
trimestre 2020

¿Por qué?



\* Twitter suma 166 millones de usuarios durante el Coronavirus. (2020, 3 junio). REBOLD, Data-Driven Marketing & Communication.

<https://letsrebold.com/es/blog/twitter-suma-166-millones-de-usuarios-frente-al-coronavirus/#:%7E:text=Asimismo%2C%20Twitter%20comunic%C3%B3%20en%20la,el%20primer%20trimestre%20de%202019.>

# 1. INTRODUCCIÓN



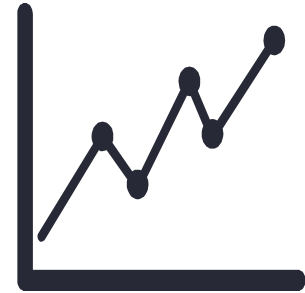
**Información en  
tiempo real**



**Diferentes  
localidades**



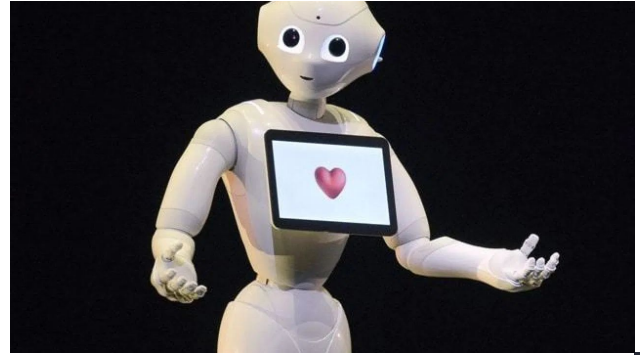
**Percepciones sobre  
temas específicos**



**Aplicación  
herramienta  
minería de datos**

# Importancia

- Demostrar que los algoritmos computacionales pueden clasificar subjetivamente al lenguaje natural.
- Demostrar que la clasificación automática puede estar muy cerca de una clasificación manual.



# 1. INTRODUCCIÓN

## Objetivos

- Identificar cómo se relaciona el sentimiento identificado con el contexto país .
- Identificar palabras que son clave a la hora de categorizar el sentimiento.
- Establecer algoritmos para predecir sentimientos de forma sistematizada.
- Entrenar modelos de clasificación en base a tweets usando un dataset de entrenamiento y un dataset de evaluación.



## 2. Hipótesis y preguntas

Objetivos del proyecto.



# Preguntas/hipotesis

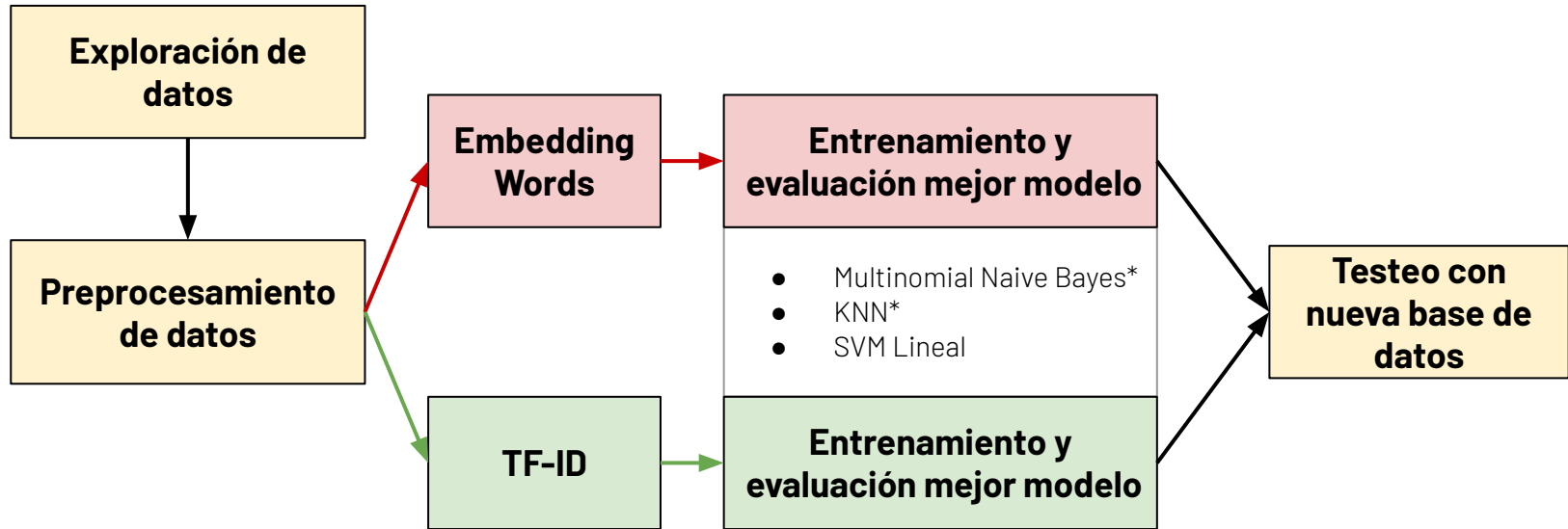
1. ¿El sentimiento general sobre el COVID-19 varía por la localidad registrada?
2. Dada las características del COVID-19 y sus consecuencias, más del 50% de los tweets están asociados a un sentimiento negativo o extremadamente negativo.
3. ¿Existe una variación en el sentimiento de los comentarios al avanzar los días? ¿Cómo se relaciona con la evolución de nuevos casos?
4. Los sentimientos de una nueva base de datos pueden ser definidos a partir de un clasificador entrenado con una base de datos previa.



# 3. Metodologia

Propuesta experimental.

### 3. Metodología



Four thick, dark blue horizontal bars stacked vertically on the left side of the slide.

# 4. Experimentos y resultados.

Desarrollando el análisis.





# 4.1. Exploración de datos

Conociendo y analizando BBDD



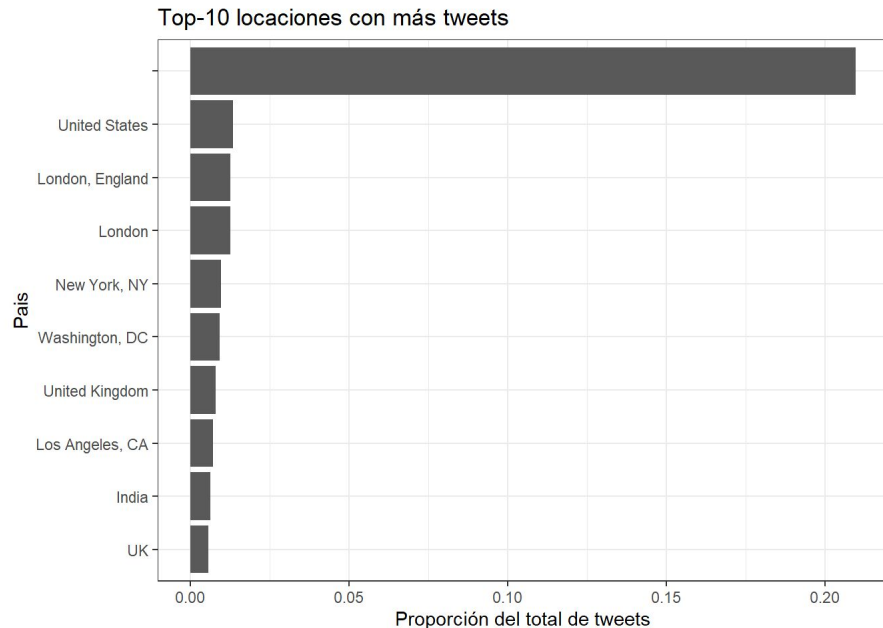
## 4.1. Descripción de la Base de datos

**44955 tweets** relacionados con el tema COVID 19.

- Atributos
  - Fecha de publicación (2 de marzo - 14 de abril)
  - Identificador de usuario
  - Localización del usuario (opcional)
    - Ciudad, País, Estado
  - Sentimiento asignado al Tweet (manual)

## 4.1. Descripción de la Base de datos

### Países y ciudades con más tweets asociados

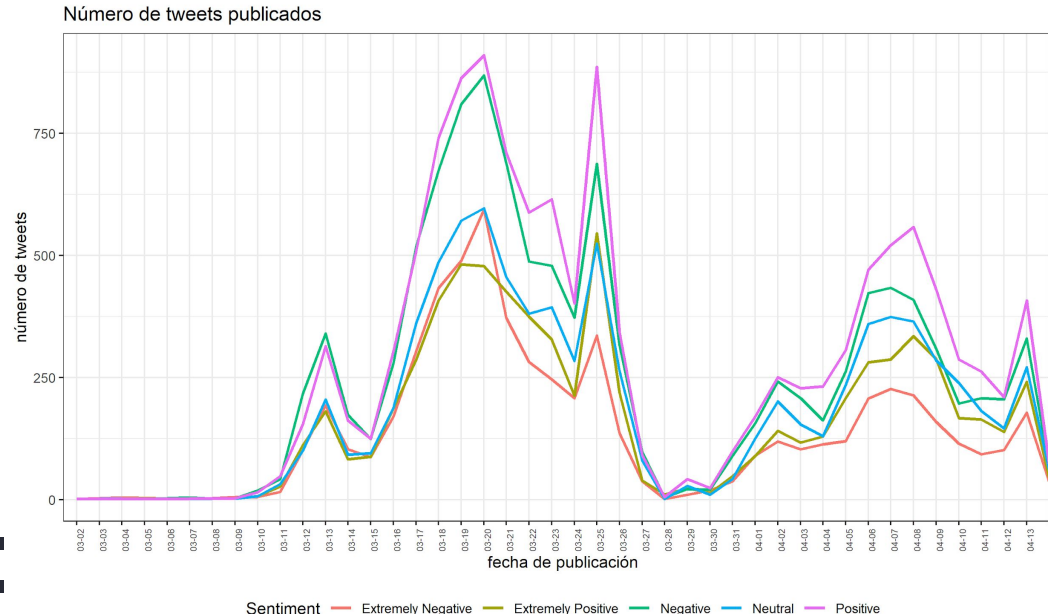


- ~20% no registra localidad.
- Datos agregados a nivel de ciudad y/o países

Resuelve P1.

## 4.1. Descripción de la Base de datos

### Evolución de sentimientos en Twitter Mar - Abr 2020



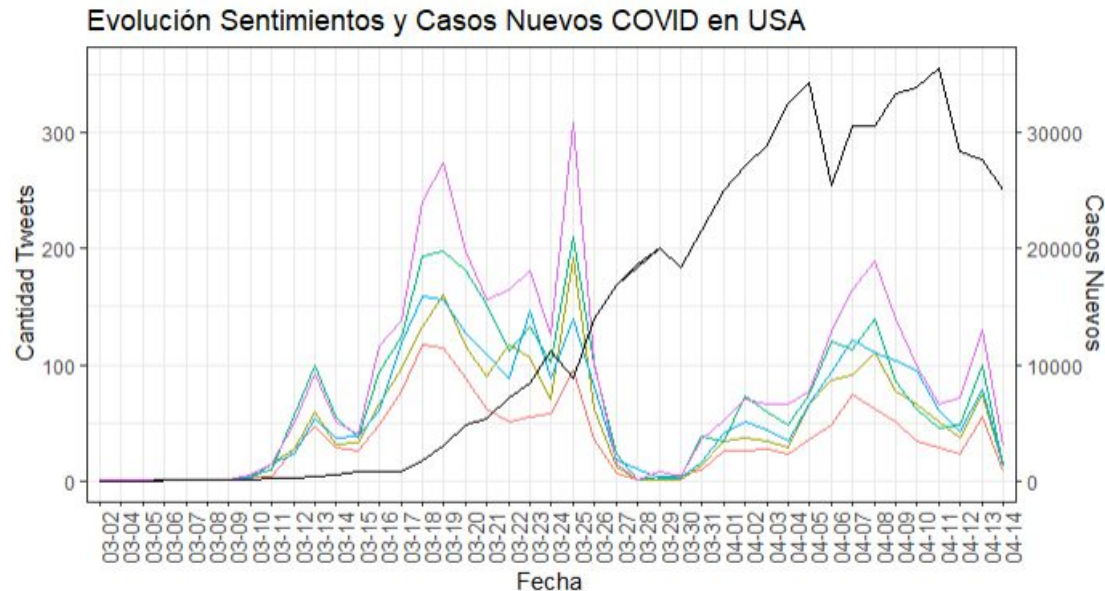
- Preponderancia “Positive” y “Negative”
- “Extremely Negative” menos común

Resuelve P2.



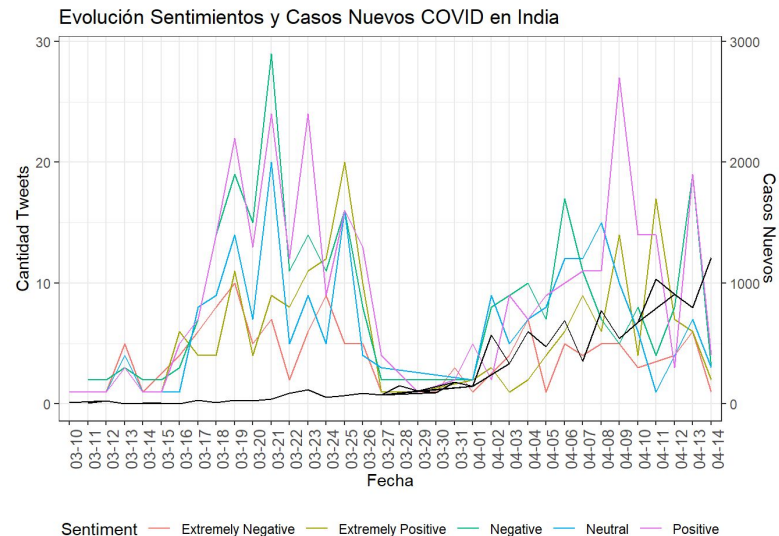
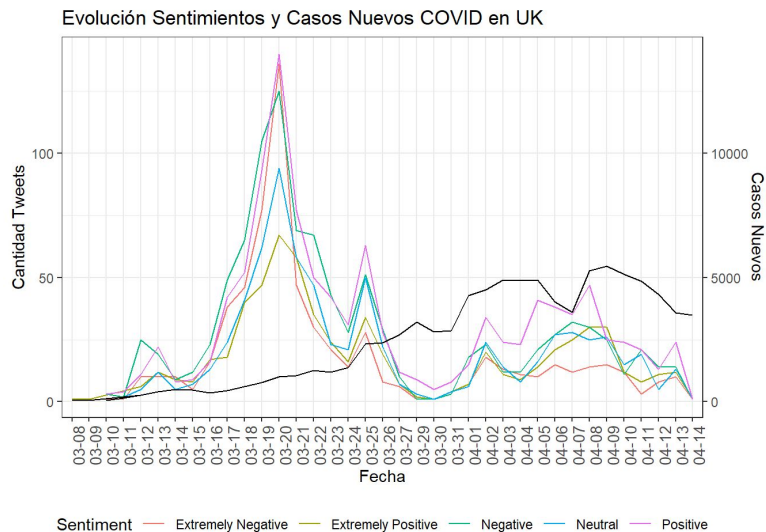
## 4.1. Descripción de la Base de datos.

### Evolución de sentimientos y casos nuevos de Covid.



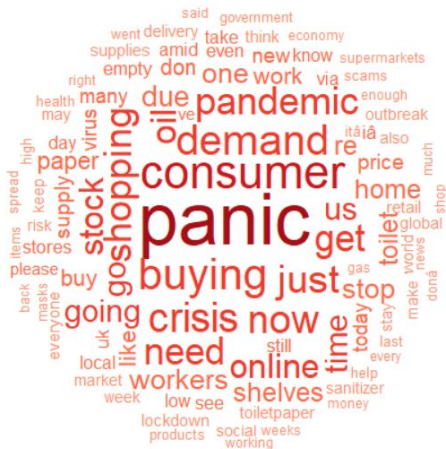
## 4.1. Descripción de la Base de datos.

### Evolución de sentimientos y casos nuevos de Covid.



## 4.1. Descripción de la Base de datos

### Palabras más usadas en Twitter sobre COVID-19 Mar - Abr 2020



**Negative + Extremely  
Negative**



**Neutral**



**Positive + Extremely  
Positive**



## 4.2. Preprocesamiento

Limpieza.



## 4.2. Preprocesamiento de datos

- Remoción de notaciones propias de Twitter (#, RT y @)
- Remoción de palabras “vacías” del lenguaje (conectores, etc.)
- Consideración de Bigramas (nombres compuestos por dos palabras)(gensim).
- Eliminación de URLs e hipervínculos (www, com, etc)



## 4.2. Vectorización

Creación de columnas



## 4.2. Vectorización

### Embedding Words (Word2Vec)

- Red neuronal con una capa oculta
- Predecir cada palabra cercana de cada término de un texto
- Obtener los pesos de la capa oculta
- Reducción de dimensionalidad: N\_tweets x 200
- **Accuracy = 0.65 (SVM)**

### TF-IDF

- TF = Frequency/total number of words in the document
- IDF =  $\log(\text{total number of documents} / (\text{Number of documents in which the word is present} + 1))$
- Medida de importancia de la palabra
- N\_tweets x 10758
- **Accuracy = 0.79 (SVM)**



## 4.4. Predicción

Aplicación del modelo al nuevo dataset.





## 4.4. Predicción

**Información del nuevo dataset.**

**179108 tweets** relacionados con el tema COVID 19.

- Atributos
  - Fecha de publicación (24 de julio - 30 de agosto)
  - Identificador de usuario
  - Localización del usuario (opcional)
    - Ciudad, País, Estado

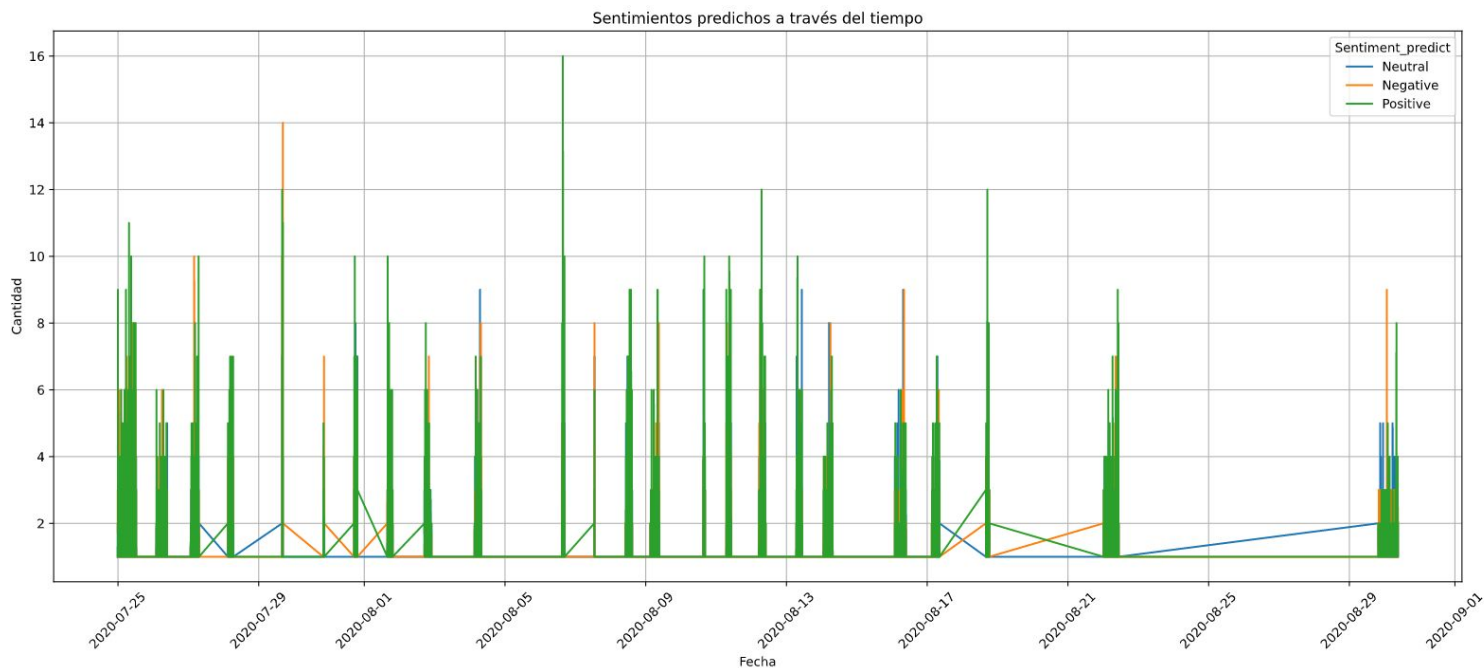
## 4.4. Predicción

### Resultados de clasificación.

Predicción de sentimiento	Cantidad de Tweets	% sobre total de Tweets
Positivo	76.488	42,7%
Negativo	55.622	31%
Neutral	46.998	26,2%

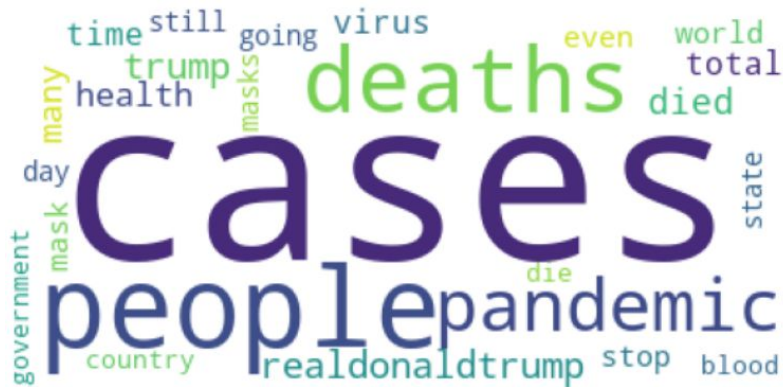
## 4.4. Predicción

### Sentimientos predichos a través del tiempo.

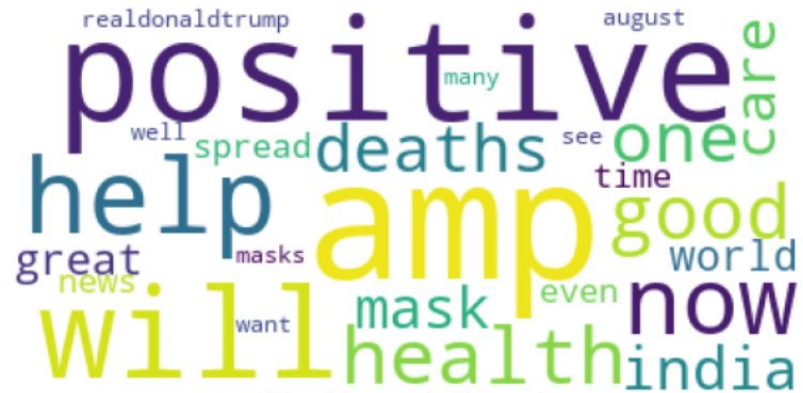


## 4.4. Predicción.

## 50 palabras más frecuentes.



## Sentimiento Negativo



## Sentimiento Positivo

## 4.4. Predicción.

### Comparación de resultados.

Predicción de sentimiento	% sobre total de Tweets (2 de marzo - 14 de abril)	% sobre total de Tweets (24 de julio - 30 de agosto)
Positivo	43.6%	42,7%
Negativo	37.9%	31%
Neutral	18.5%	26,2%

Considerar poca representatividad de los tweets

- Aparente disminución de comentarios negativos y aumento de comentarios neutrales.



## 5. Direcciones futuras

## 5. Direcciones futuras.

- Investigar más métodos para clasificar el lenguaje natural y comparar sus métricas con las métricas de los métodos enseñados en el curso.

