Detección de sarcasmo

Micaela Rubin
Pablo Torres

¿Eso es sarcasmo?

- El **sarcasmo** es una forma del lenguaje en la que los individuos declaran lo contrario de lo que está afirmando.
- Con esta ambigüedad intencional, la detección del sarcasmo (**DS**) siempre es una tarea desafiante, incluso para los humanos.
- En ciencias de datos y computación, esta tarea pertenece esencialmente al dominio del **procesamiento del lenguaje natural**.

Procesamiento del lenguaje natural (PLN o NLP)

- El **procesamiento del lenguaje natural** —en inglés, *natural language processing*, NLP— es un campo de las ciencias de la computación, de la inteligencia artificial y de la lingüística que estudia las interacciones entre las computadoras y el lenguaje humano.

- Algunas de sus principales aplicaciones incluyen síntesis del discurso, análisis del lenguaje, comprensión del lenguaje, reconocimiento del habla, síntesis de voz, generación de lenguajes naturales y traducción automática.

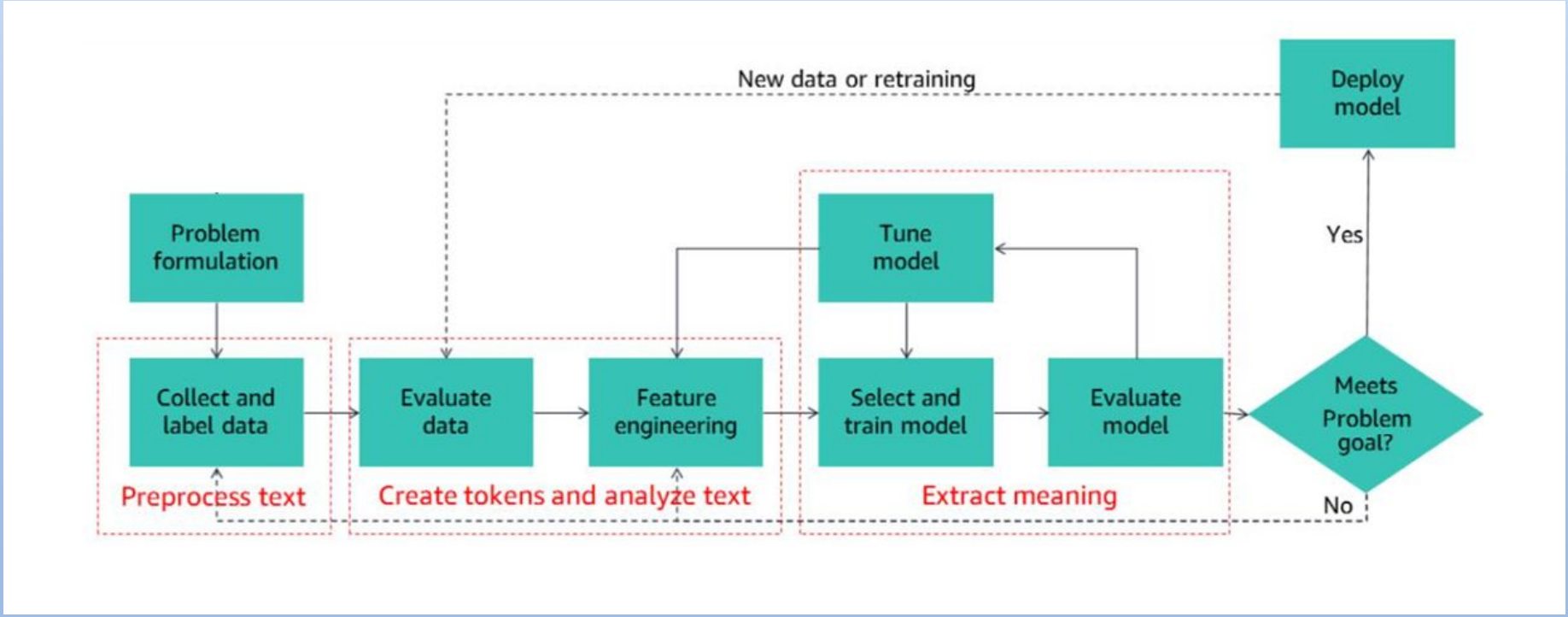
Análisis de Sentimiento

- En términos generales, el análisis de sentimiento intenta determinar la actitud de un interlocutor o usuario con respecto a algún tema o la polaridad contextual general de un documento.

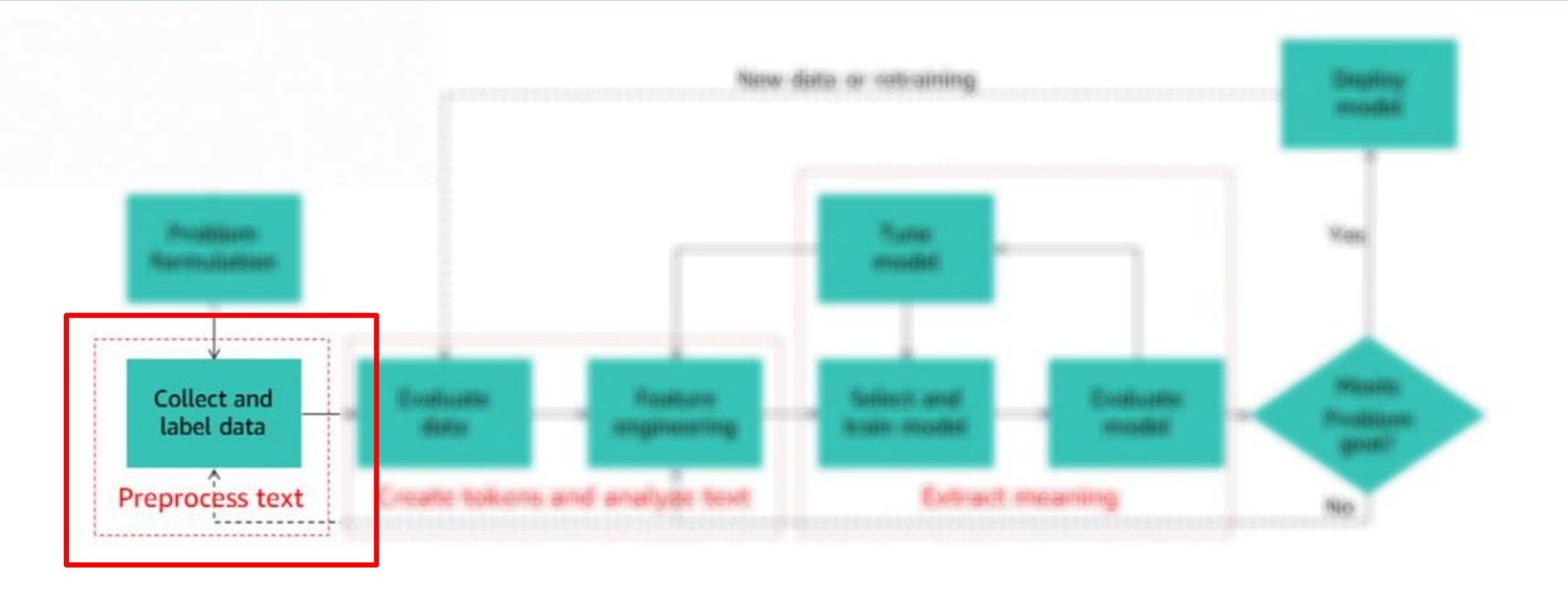
- Formalmente podríamos definir la detección de sarcasmo como:

> Dado un comentario *t* sin etiquetar en un conjunto de comentarios **U**, una solución para la detección del sarcasmo tiene como objetivo detectar automática y claramente si es sarcástico *t* o no.

Flujo de trabajo del proyecto de DS



Presentación del corpus



Nuestro dataset

- Un corpora compuesto por 3 corpus de comentarios que fueron realizados y que se encontraron en la web.

- Los documentos de cada uno de ellos poseen diferencias en su estructura, uno de ellos captura hipérboles (HYP), otro hace referencia a preguntas retóricas (RQ), y el último corpus captura hipérboles, preguntas retóricas y otros tipos de sarcasmo (GEN).

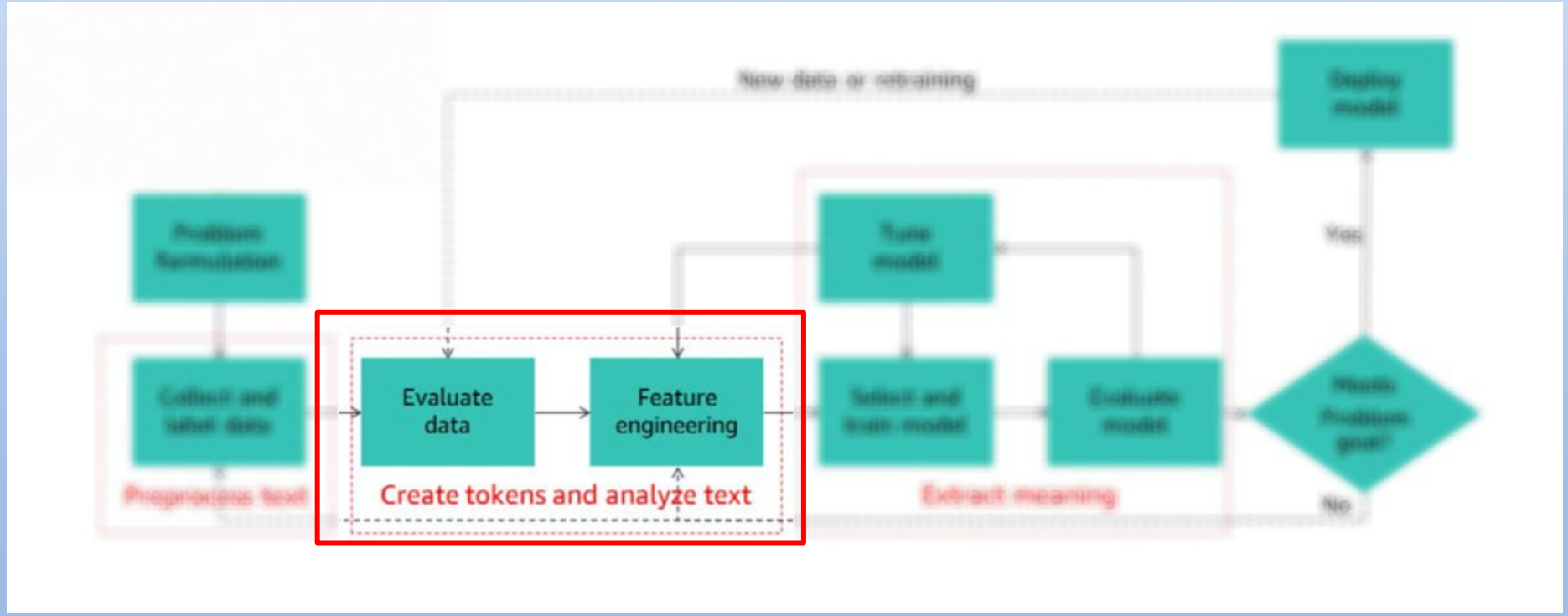
Descripción del corpora

- Los 3 corpus se encuentran organizados por: clase, id y text. La clase hace referencia a la etiqueta de los documentos “sarcasmo” y “no sarcasmo”. El id es un número de identificación del documento y test es el comentario propiamente dicho.

- El corpora entero está compuesto por 9386 documentos de los cuales 6520 pertenecen al corpus GEN, 1702 al corpus de RQ y 1164 al de HYP.

¿Es esta información relevante? ¿Es de ayuda que el corpora se divida en corpus con diferentes tipos de sarcasmo? ¿Utilizamos el corpora como un solo dataset o analizamos los corpus por separado?

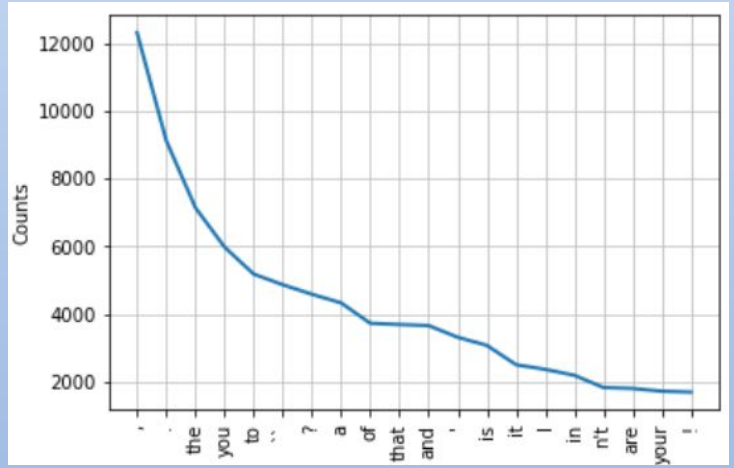
Evaluación de datos y Feature Engineering



Evaluación de los datasets

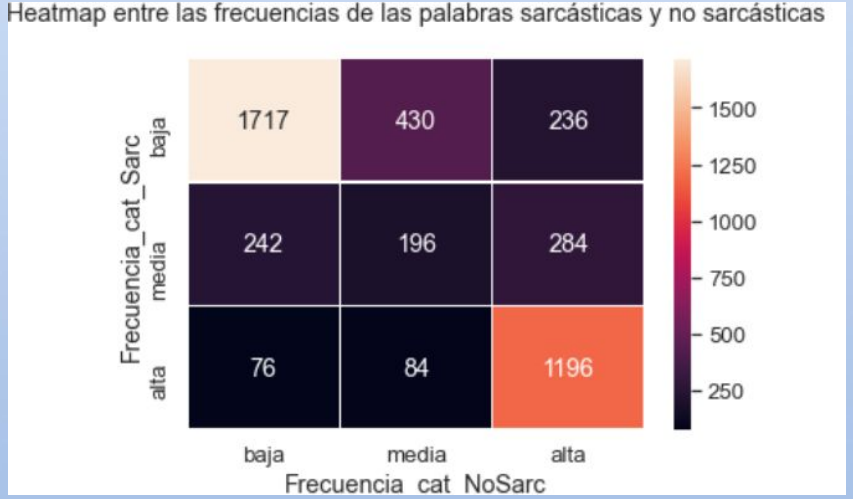
- En una primera instancia tokenizamos con el tokenizer de la librería NLTK y analizamos las cuáles eran las palabras más frecuentes.

  Resultados relevantes: a priori en el dataset GEN no había diferencias fundamentales entre comentarios SARC y no SARC.



- Luego generamos modelos de bigramas e hicimos lo mismo utilizando NLTK. Obtuvimos resultados similares a unigramas (excepto para el corpus RQ, que poseía una cantidad elevada de signos de interrogación).

  Frecuencia de palabras en el corpus GEN: sorprendió la proporción de frecuencia media de no SARC que son bajas en SARC



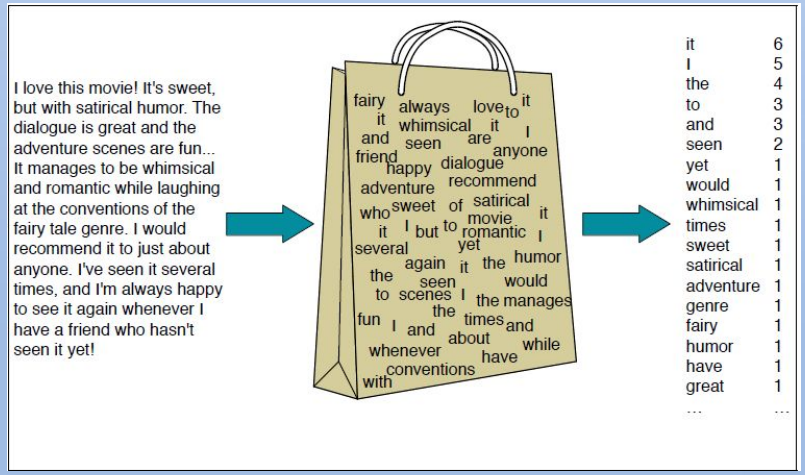
- Analizamos las entidades más importantes en el corpus con la librería **spaCy**. No hubo datos reveladores.

Feature engineering

Dos tipos distintos de procesamientos sobre los dataset:

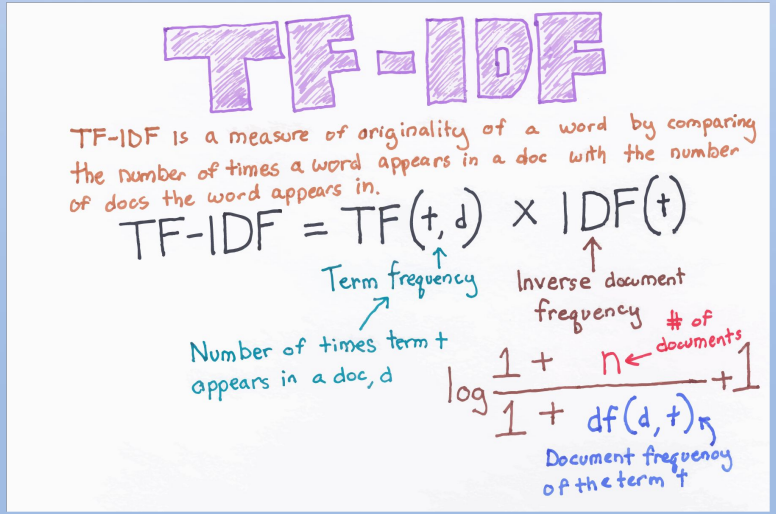
- **Bag of words**: transforma a cada documento del corpus un conjunto desordenado de palabras. La posición de cada palabra en el documento original es ignorada; y se conserva solo su frecuencia en el documento. El resultado es un vector con n features, siendo n la cantidad de palabras del documento.

Utilizamos **CountVectorizer de Scikit Learn** tomando ngramas de grado 1 a 3.

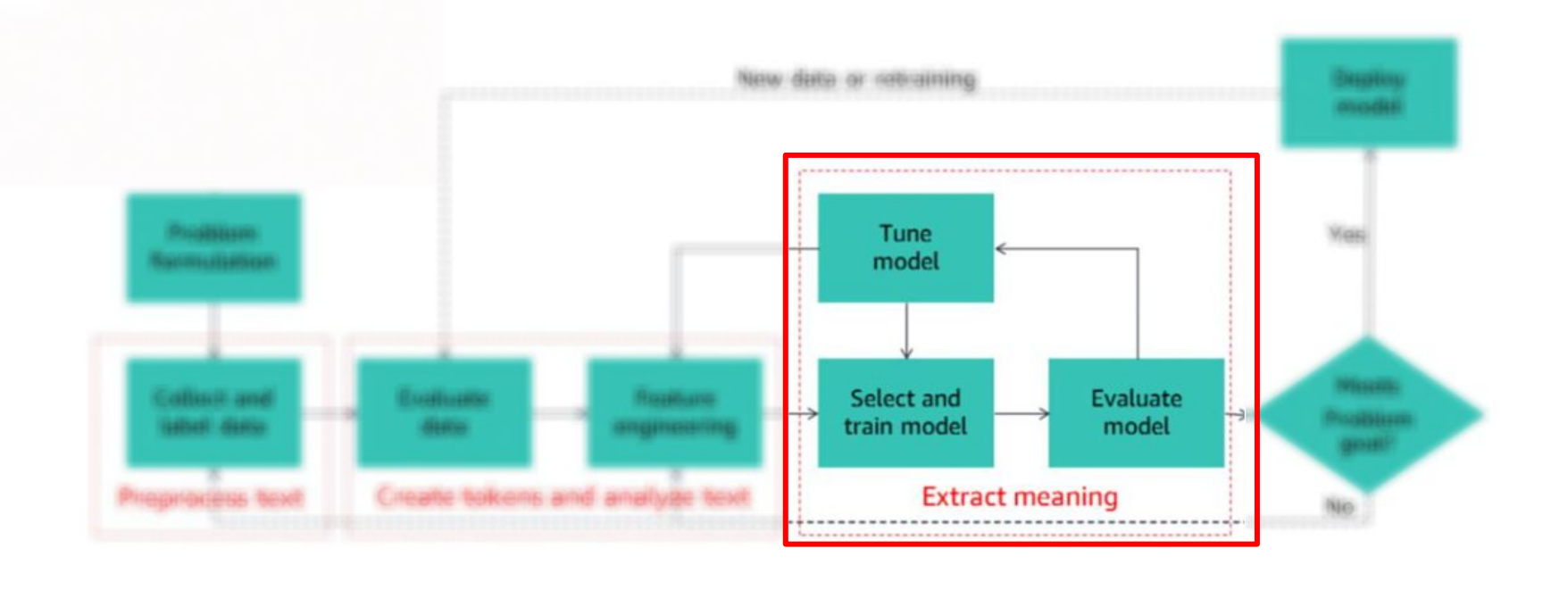


- **TF-IDF:** del inglés *Term frequency – Inverse document frequency*, frecuencia de término – frecuencia inversa de documento,es una medida numérica que expresa cuán relevante es una palabra para un documento en una colección. El valor tf-idf aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero es compensada por la frecuencia de la palabra en la colección de documentos.

  Utilizamos **TfidfVectorizer de Scikit Learn**, tomando ngramas de grado 1 a 3.



Entrenamiento de modelos



Regresión logística

Entrenado con los corpus procesados con **Bag of Words**:

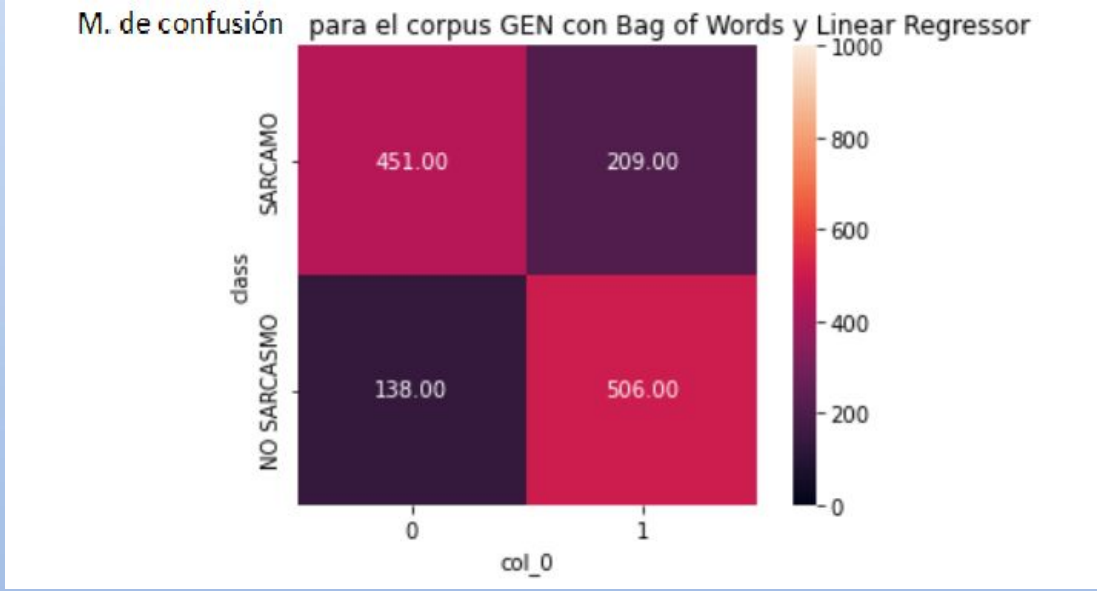
	F1 Score	Accuracy Score	Precission	Recall
GEN	0.744665	0.733896	0.707692	0.785714
HYP	0.649351	0.652361	0.619835	0.681818
RQ	0.714286	0.706745	0.672043	0.762195

Entrenado con los corpus procesados con **TF-IDF**:

	F1 Score	Accuracy Score	Precission	Recall
GEN	0.739130	0.733129	0.714493	0.765528
HYP	0.628319	0.639485	0.612069	0.645455
RQ	0.722892	0.730205	0.714286	0.731707

Se puede ver que en ambos casos el corpus que mejor scores obtuvo es el GEN.

Ejemplo de matriz de confusión para el regresor logístico GEN con Bag of Words



Se obtuvieron más del doble de verdaderos positivos y negativos que de falsos positivos y negativos, reflejando que los modelos alcanzaron scores por arriba de un 60 %.

Naive Bayes

Entrenado con los corpus procesados con **Bag of Words**:

	F1 Score	Accuracy Score	Precission	Recall
GEN	0.736156	0.751534	0.773973	0.701863
HYP	0.628099	0.613734	0.575758	0.690909
RQ	0.662420	0.689150	0.693333	0.634146

Se puede ver que el corpus genérico es el que obtuvo los mejores scores nuevamente.

Entrenado con los corpus procesados con **TF-IDF**:

	F1 Score	Accuracy Score	Precission	Recall
GEN	0.643068	0.721626	0.876676	0.507764
HYP	0.653696	0.618026	0.571429	0.763636
RQ	0.643599	0.697947	0.744000	0.567073

Tanto para el caso del corpus genérico como el de preguntas retóricas presentan un bajo recall y una precisión alta, pero con el corpus de hipérboles sucede lo contrario.

Regresión lineal de Bag of Words con SVD

	F1 Score	Accuracy Score	Precission	Recall
GEN	0.721783	0.703221	0.672021	0.779503
HYP	0.621277	0.618026	0.584000	0.663636
RQ	0.651163	0.648094	0.622222	0.682927

- Los resultados obtenidos no tuvieron una mejora sustancial con respecto a los modelos anteriores, y en algunos casos los scores descendieron.

- Al utilizar los corpus de preguntas retóricas e hipérboles, no se obtienen mejores scores que al utilizar el corpus genérico.

Conclusiones parciales

Nuestro trabajo está lejos de concluir, sin embargo, podemos sacar algunas **conclusiones provisorias** con los resultados actuales:

- Los distintos corpus arrojaron performance diferentes.
- El tamaño de los corpus puede influir en la performance.
- Los modelos (Bag of Words, TFIDF) de preprocesamiento impactan de forma distinta en el entrenamiento con Naive Bayes y Regresión Logística.
- Queda por ver qué sucedería con otro tipos de embeddings y reducción dimensional
- Considerando el F1 score y la precisión y la sensibilidad actual, los modelos tienen una perfomance que deja mucho que desear.

Consideraciones y mejoras

Estaría bueno que pueda haber una materia optativa sobre procesamiento del lenguaje natural, o que en las materias obligatorias haya ejemplos aplicados a NLP ya que es difícil poder plasmar los conceptos aprendidos en las materias obligatorias, a una mentoría sobre NLP.


SARCASM
DETECTOR