

Proyecto Final Bootcamp Análisis de Datos

Upgrate Hub

Alumno: Osvaldo González Prieto

Noviembre - 2024





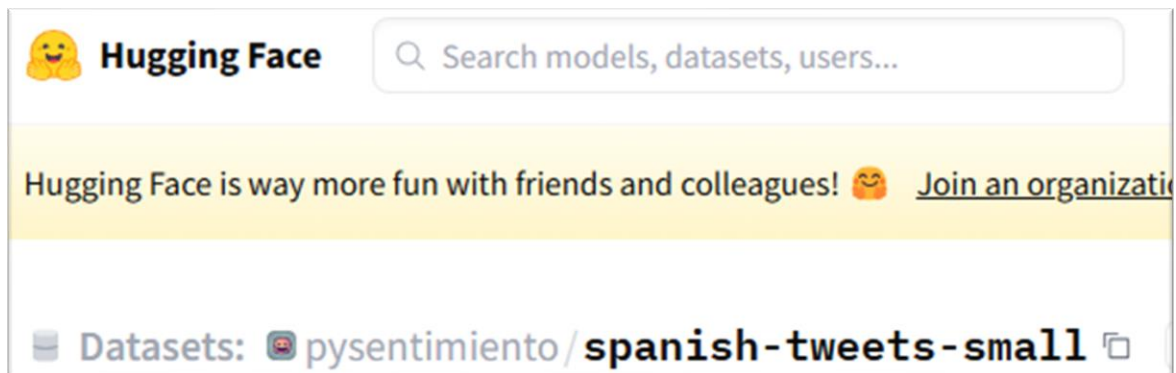
Lenguaje de Procesamiento Natural

Objetivos

- Descubrir las posibilidades que brinda el LPN
- Aprender a utilizar las diferentes herramientas.
- Conocer las diferentes posibilidades de presentar los resultados.

Fuente Dataset

<https://huggingface.co/datasets/pysentimiento/spanish-tweets-small>



Homepage
github.com

Paper
RoBERTuito: a pre-trained language model for social m...

Point of Contact:
[jmperez \(at\) dc.uba.ar](mailto:jmperez@dc.uba.ar)

Size of downloaded dataset files:
2.63 GB

Size of the auto-converted Parquet files:
2.63 GB

Number of rows:
31,123,665

Fuente Dataset

—

Características

Dataset Viewer

Auto-converted to Parquet

API




Embed

Full Screen Viewer

Split (2)
train · 24.9M rows

Search this dataset

SQL Console

text string · lengths	tweet_id string · lengths	user_id string · lengths
		
@Liz_Mile Y después dices que no eres fan... https://t.co/J91jKvugkq	1272762928475918336	3247162977
@Liz_Mile @Yaniserrano @ChasKapop Jajajaja si se deja crecer el cabello estaría bien 🍷 me gusta la idea pero es...	1272762440556654592	3247162977
Hay que tener pulso nivel 500 para darle con exactitud, parece fácil pero no 😊 lo experimente cuando el engreído d...	1272736704257052673	3247162977
@Liz_Mile Otro chino mas?!!!!!! https://t.co/LsEBYDLcM3	1272734775720267776	3247162977
@Yaniserrano Más extensiones 🤔	1272734480231469056	3247162977
@ChasKapop Calidad XD 😊	1272734289298370561	3247162977
@Yaniserrano X2 KHJ nos debe un papel de villano 🗨️	1272633266659622918	3247162977

< Previous 1 2 3 ... 248,990 Next >

Preprocesamiento del Texto

Antes de aplicar cualquier técnica de PLN, es fundamental realizar un preprocesamiento del texto para limpiarlo y estructurarlo.

Limpieza básica: Elimina URLs, menciones (@usuario), hashtags (#hashtag), emojis y caracteres especiales.

Normalización: Convierte el texto a minúsculas.

Tokenización: Divide el texto en palabras o tokens usando bibliotecas nltk y spaCy.

Stopwords: Elimina palabras comunes como "el", "la", "y", etc., que no aportan mucho significado, usando listas predefinidas como en `nltk.corpus.stopwords`.

Nube de Palabras más utilizadas

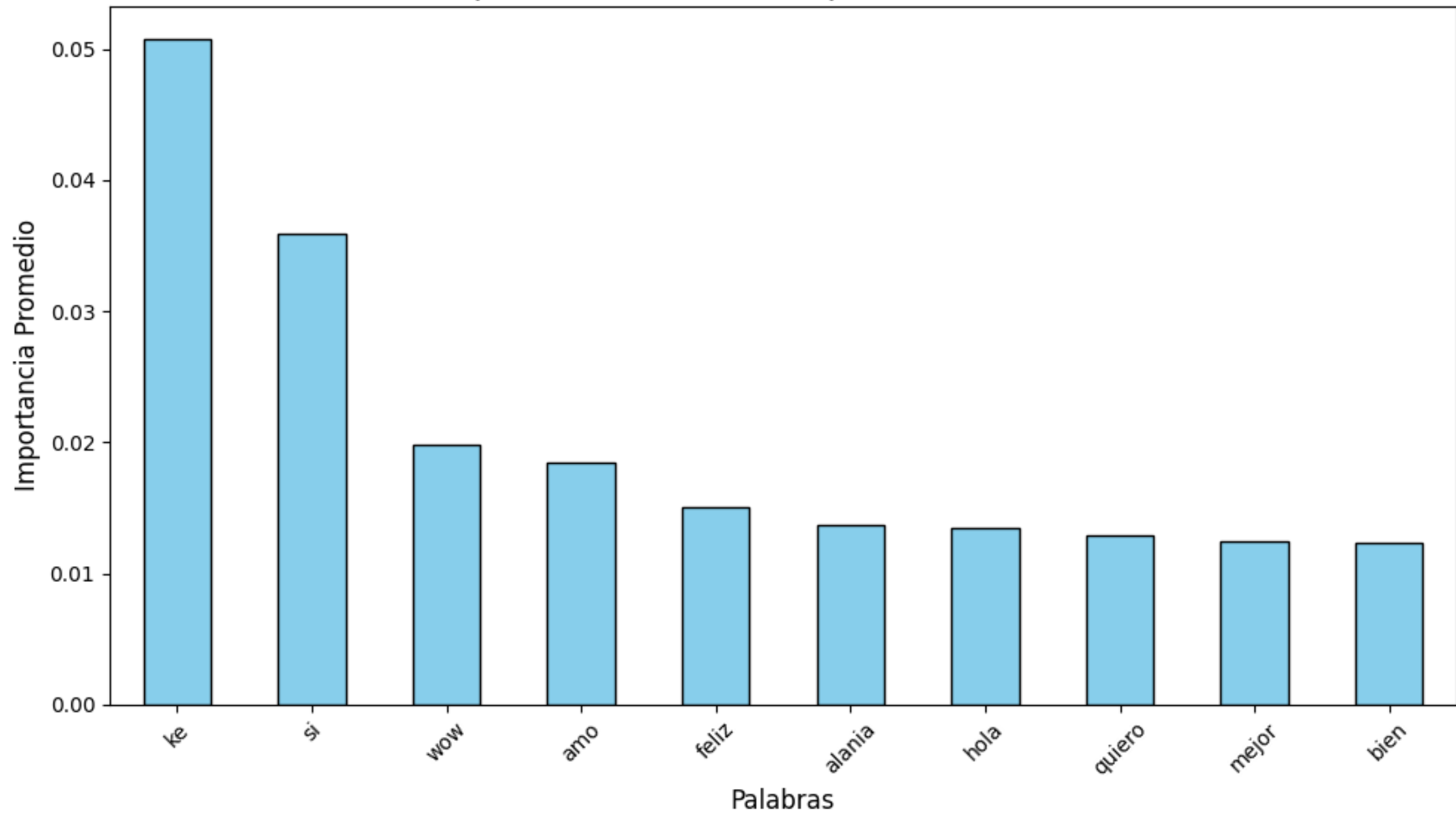


Nube de Palabras Basada en TF-IDF

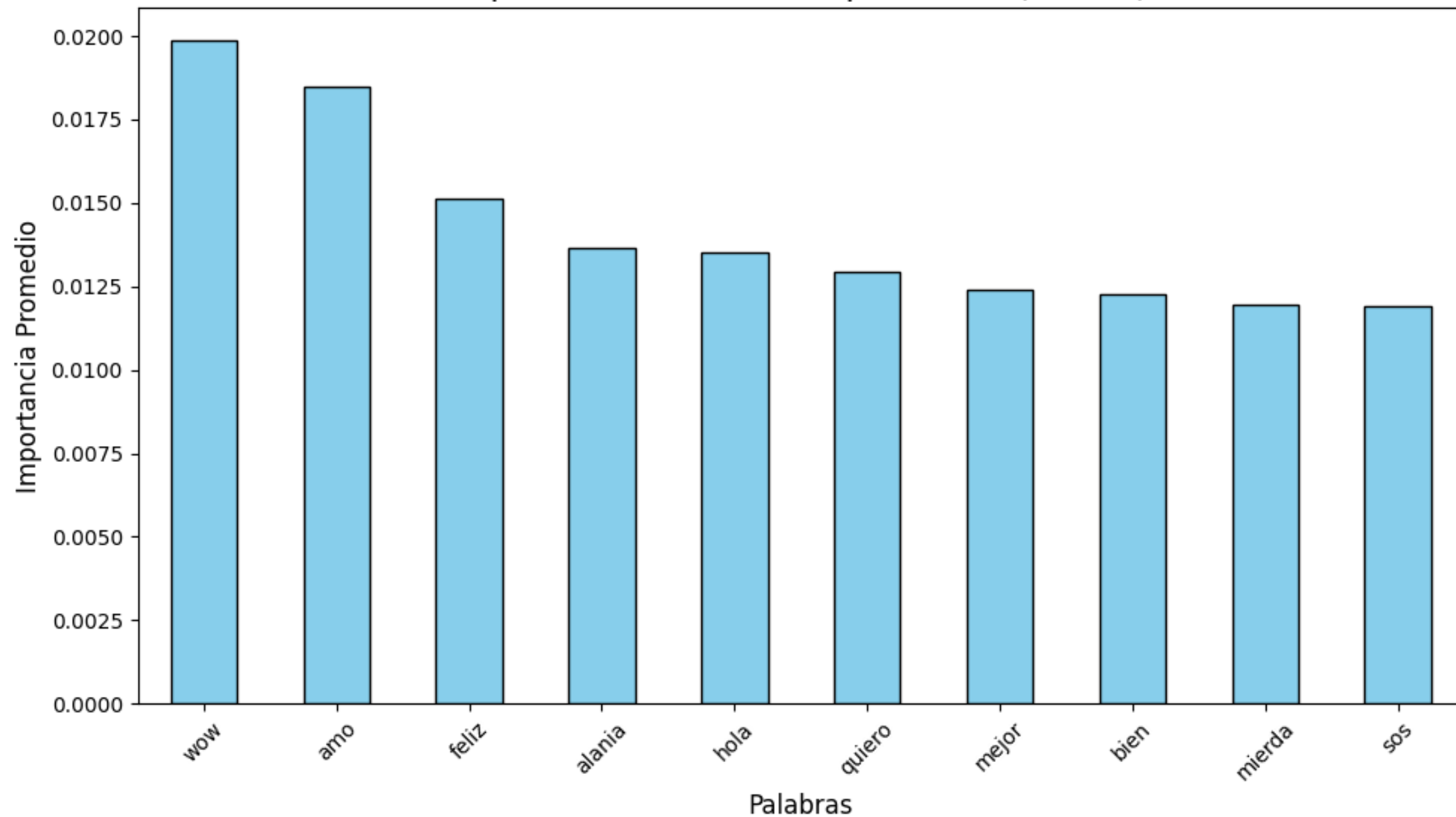


TF-IDF mide la relevancia de una palabra en un documento dentro de un conjunto. Valora términos frecuentes en el documento pero menos comunes globalmente.

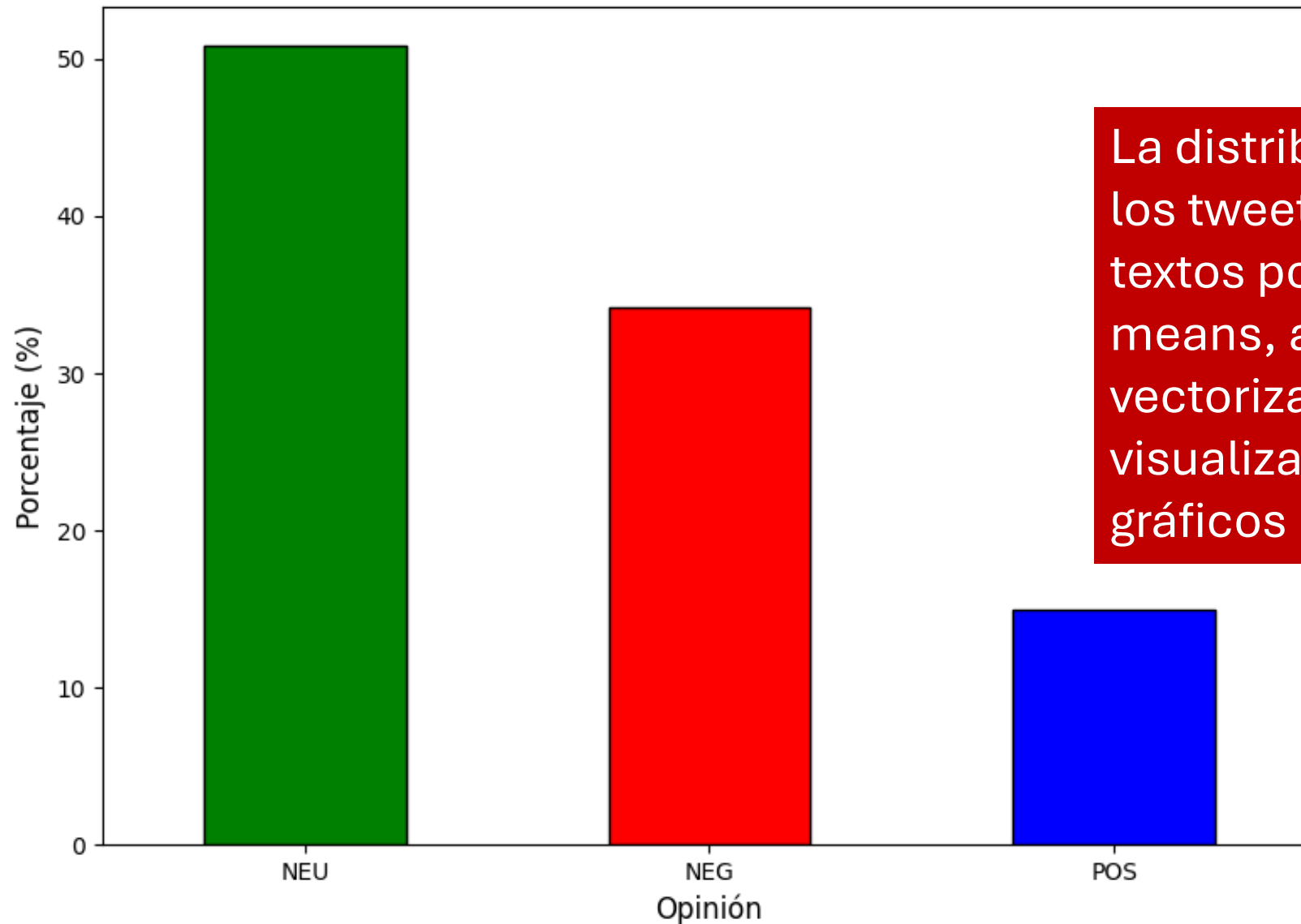
Top 10 Palabras Más Importantes (TF-IDF)



Top 10 Palabras Más Importantes (TF-IDF)

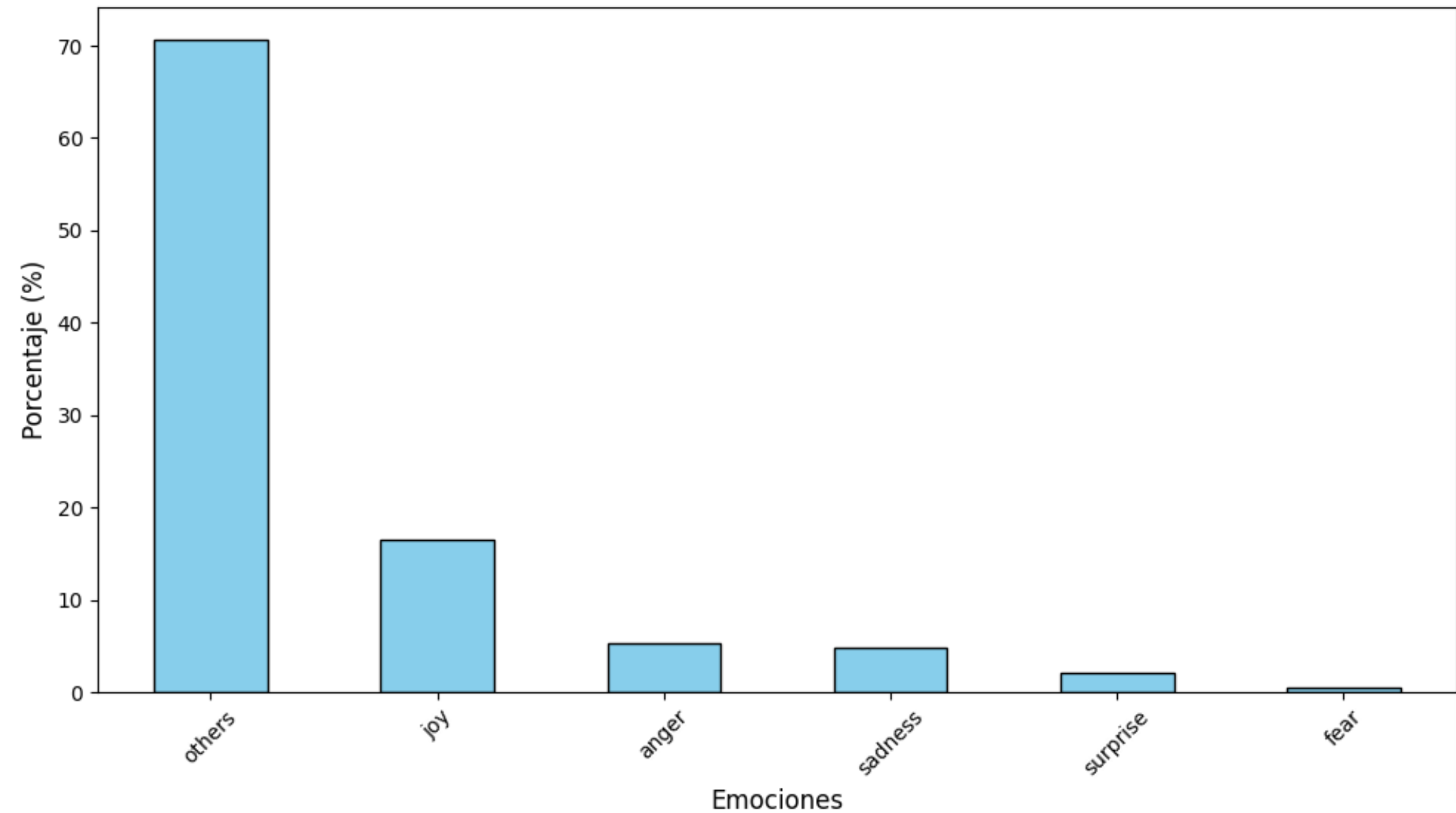


Distribución de Opiniones en los Tweets

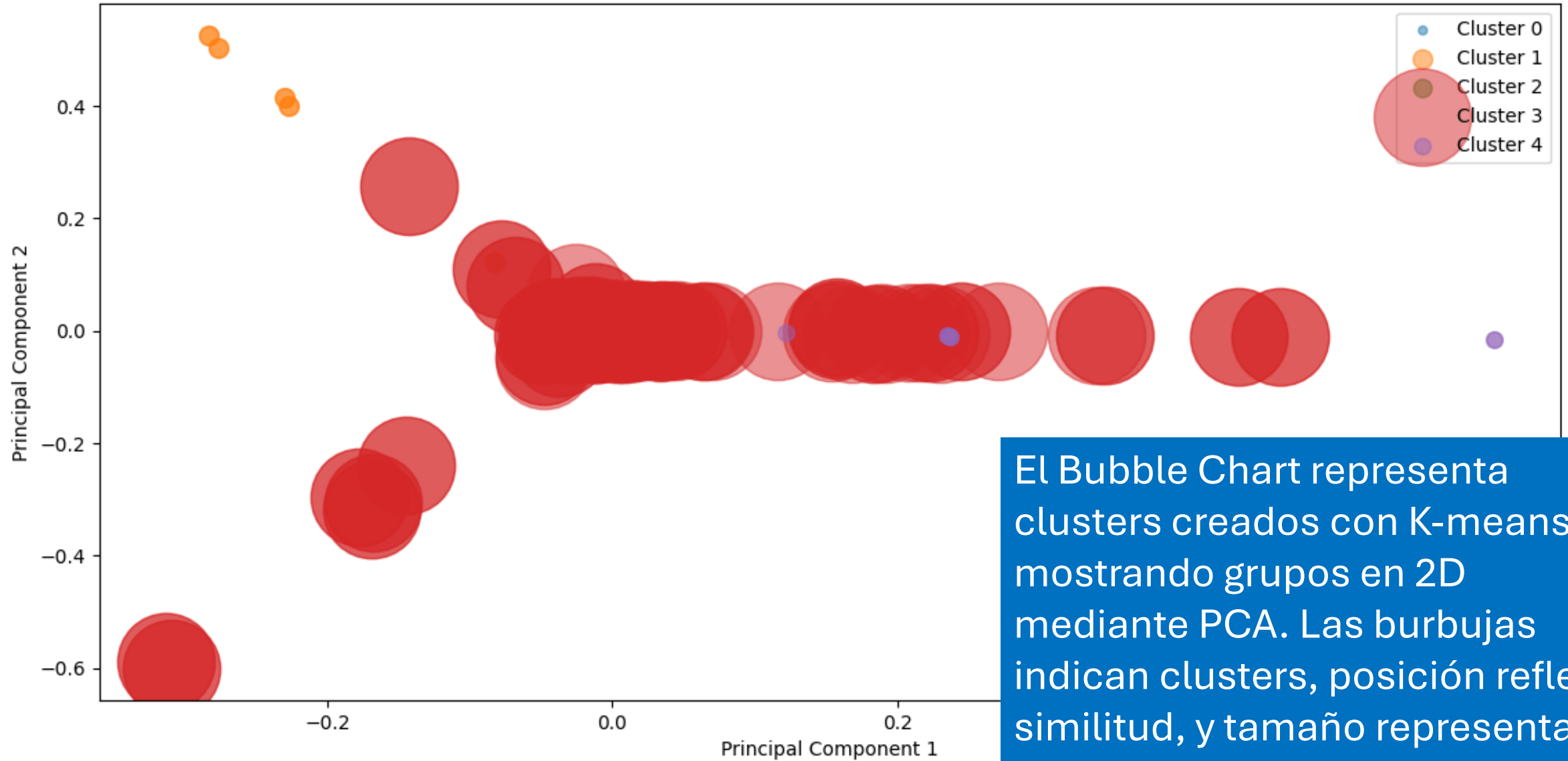


La distribución de opiniones en los tweets se realizó agrupando textos por similitud utilizando K-means, aplicando TF-IDF para vectorizar palabras clave y visualizando los clusters con gráficos basados en análisis PCA.

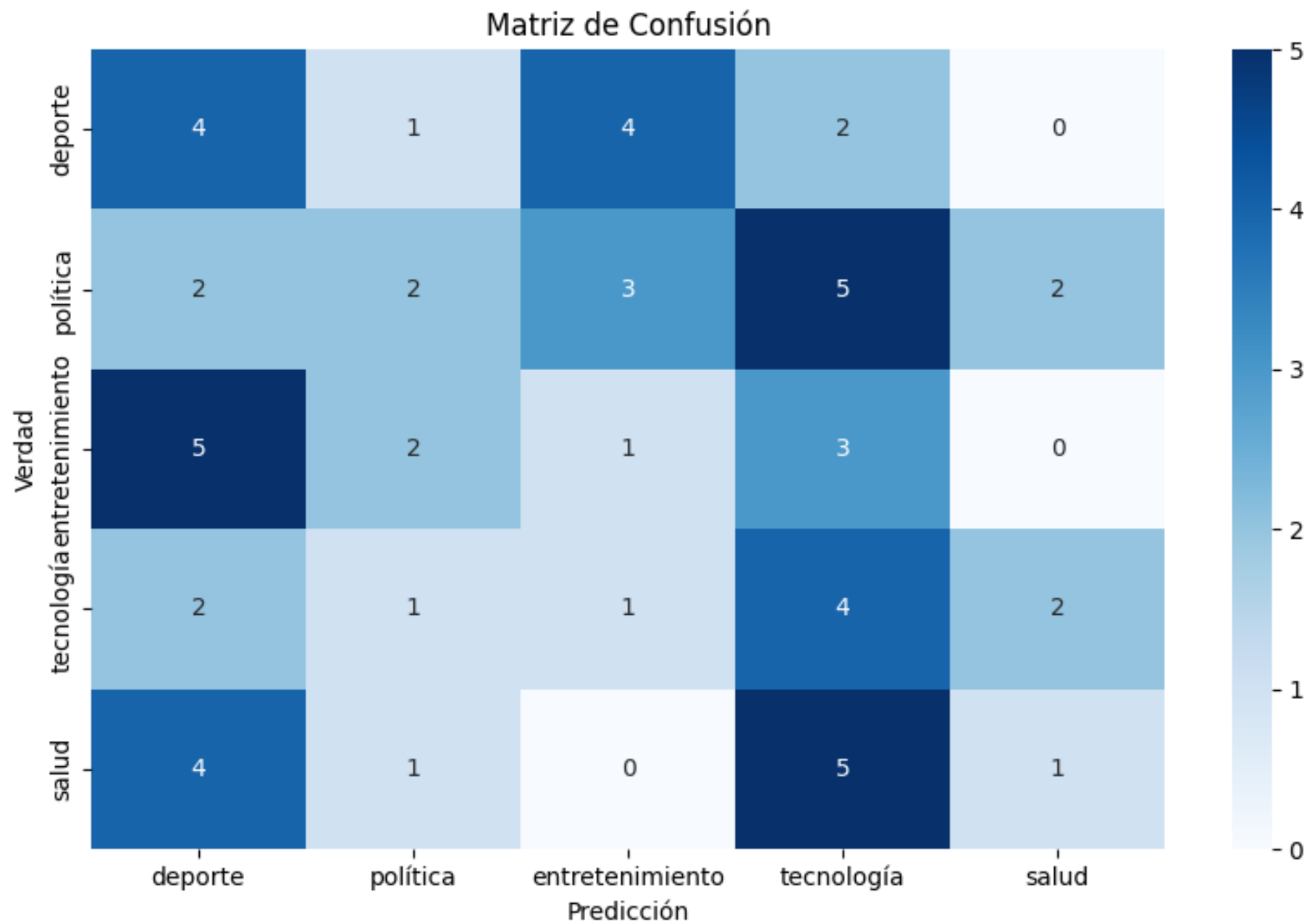
Distribución de Emociones en los Tweets



Bubble Chart of Clusters (K-means)



El Bubble Chart representa clusters creados con K-means, mostrando grupos en 2D mediante PCA. Las burbujas indican clusters, posición refleja similitud, y tamaño representa cantidad de tweets en cada grupo.



Generar tweet nuevos – con el dataset y GPT

- *Hoy es un buen día para empezar algo nuevo.*
- *Qué partidazo anoche Este equipo nunca deja de sorprenderme.*
- *La película que vi ayer me dejó sin palabras, recomendada.*

