Detección de lenguaje en textos

IA Nexus

Integrantes
Ricardo Bernabé Nicolás
Carlos Eduardo González Arceo

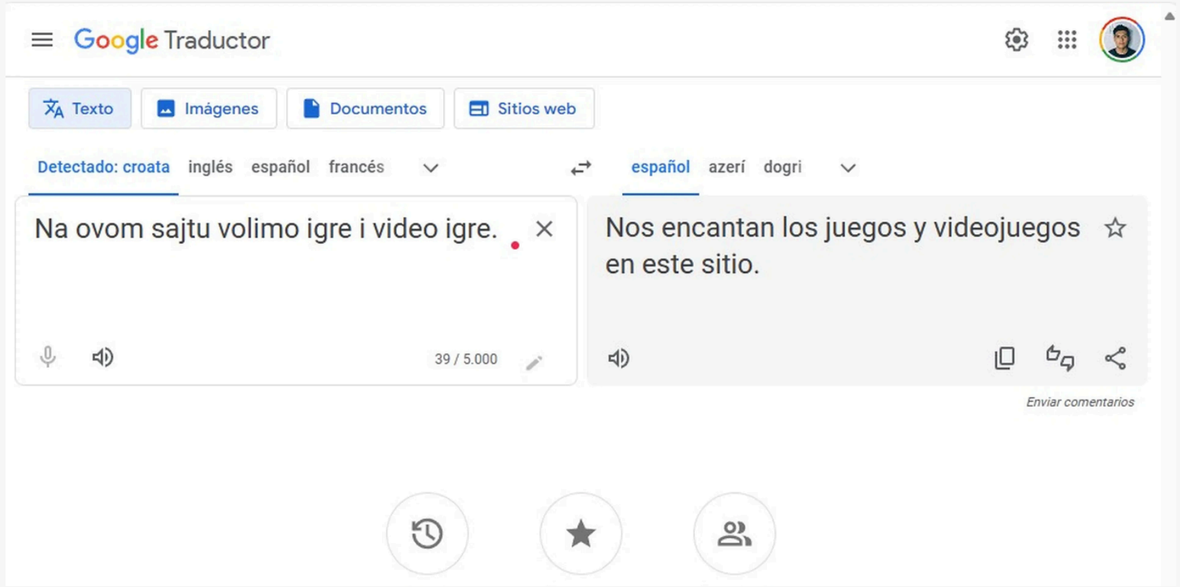
CONTENIDO

Motivación

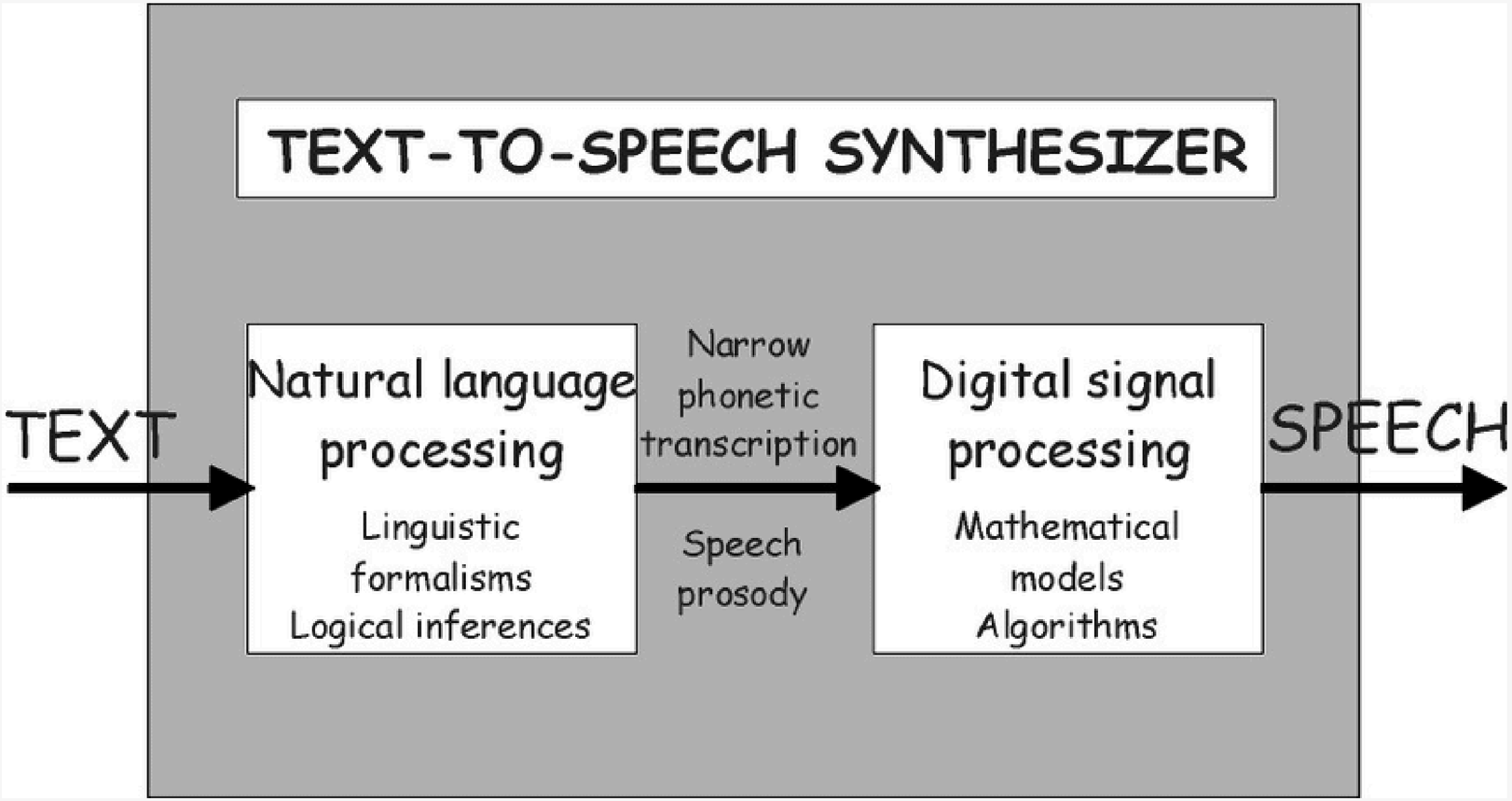
La detección automática del idioma de un texto es fundamental para aplicaciones de procesamiento de lenguaje natural (NLP), como traducción automática, análisis de sentimientos y clasificación de documentos. Ejemplos de uso: Google Traductor, herramientas de Speech to Text y Text to Speech.

Ejemplo: Función detectar idioma

Si queremos hacer una traducción de un texto de un lenguaje que desconocemos, podemos utilizar esta función para detectar el idioma y después traducir



Ejemplo: Speech To text - Text to speech



Dataset

El dataset consiste en 3698 textos con su respectiva etiqueta del idioma.

1000 Español

1000 Inglés

1000 Francés

698 Italiano

	Text	language	code
0	en navidad de poco después de que interpretó ...	Spanish	0
1	según el censo de [] había personas residien...	Spanish	0
2	en la copa mundial de fútbol sub- de pitó los...	Spanish	0
3	ally y buttons encuentran el descodificador y ...	Spanish	0
4	los primeros habitantes se establecieron cerca...	Spanish	0
...	...	...	...
995	on march empty mirrors press published epste...	English	3
996	he [musk] wants to go to mars to back up human...	English	3
997	overall the male is black above and white belo...	English	3
998	tim reynolds born december in wiesbaden germ...	English	3
999	the total high school population was now appro...	English	3

3698 rows × 3 columns

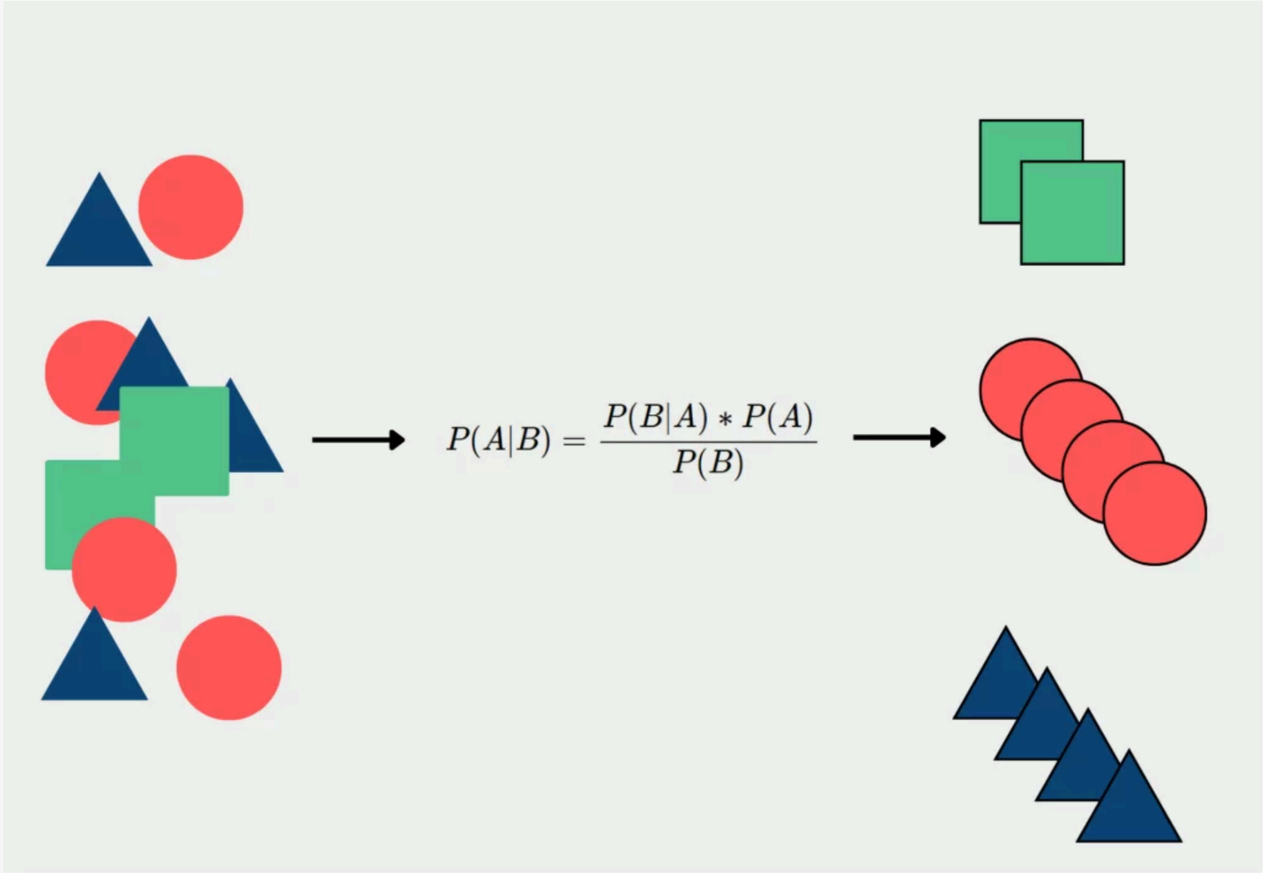
Generación de vectores

Se utilizó primero countVectorizer para contar las apariciones de cada token y después se utilizó TfidfTransformer para generar la matriz de pesos de tf-Idf

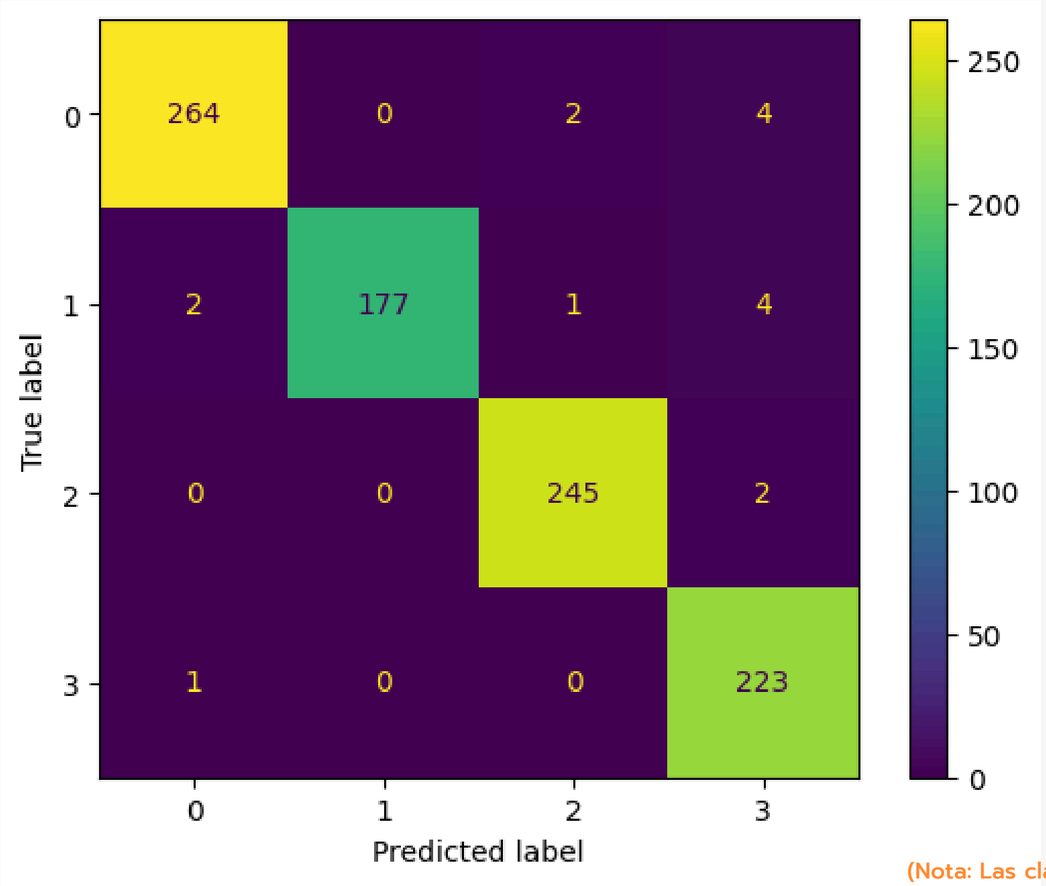
	0	1	2	3	4	5	6	7	8	9	...	35778	35779	35780	35781	35782	35783	35784	35785	35786	35787
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	2	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Naive Bayes

Naive Bayes se basa en calcular la probabilidad de que un elemento pertenece a una clase, dada la descripción del elemento, en este caso, el producto de la probabilidad de sus tokens dado la probabilidad de la clase



Resultados



(Nota: Las clases son 0: "Spanish", 1: "Italian", 2: "French", 3: "English")

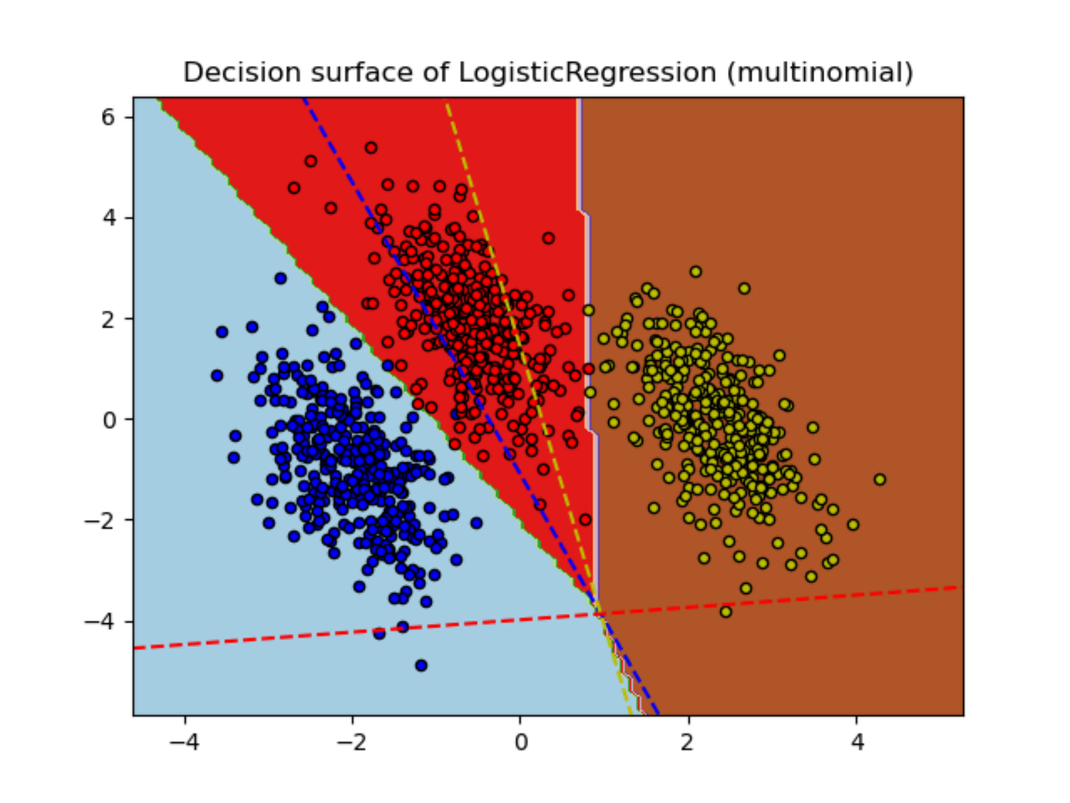
Resultados

	precision	recall	f1-score	support
0	0.99	0.98	0.98	270
1	1.00	0.96	0.98	184
2	0.99	0.99	0.99	247
3	0.96	1.00	0.98	224
accuracy			0.98	925
macro avg	0.98	0.98	0.98	925
weighted avg	0.98	0.98	0.98	925

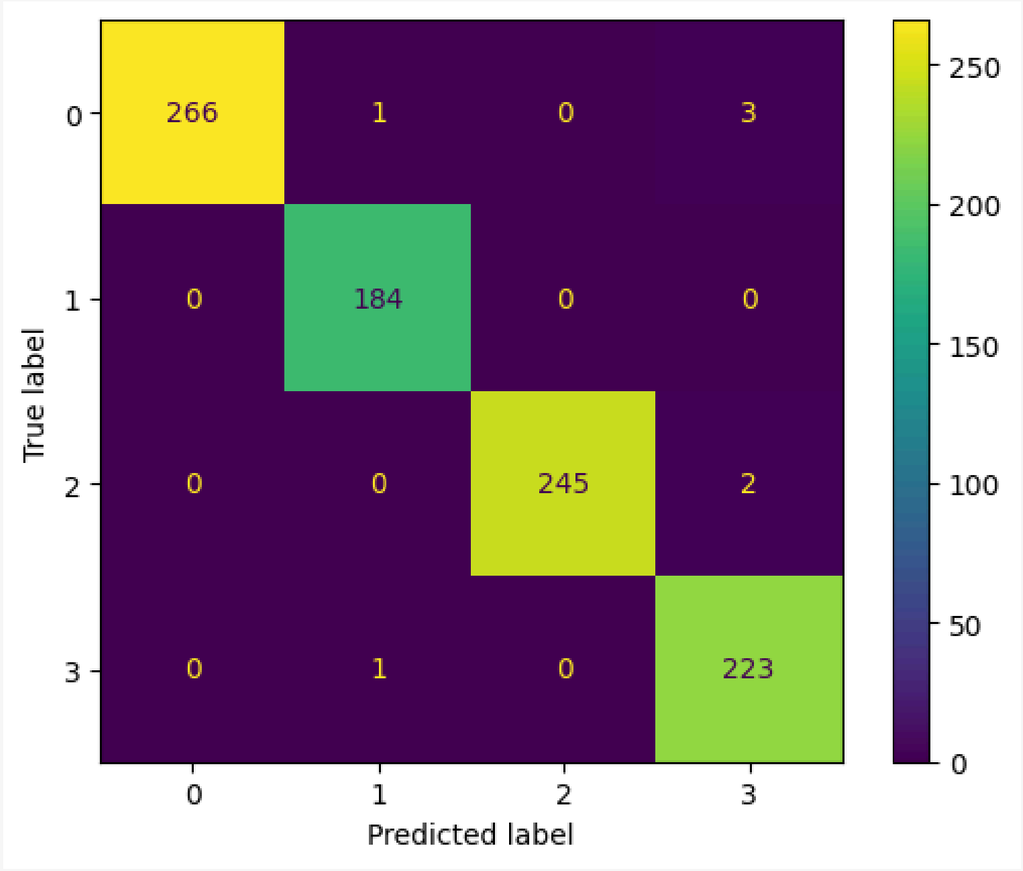
(Nota: Las clases son 0: "Spanish", 1: "Italian", 2: "French", 3: "English")

Logistic Regression Multinomial

Utiliza la creación de varios modelos 1 vs todos para obtener varias probabilidades y poder elegir la probabilidad más alta, que define a la clase



Resultados



(Nota: Las clases son 0: "Spanish", 1: "Italian", 2: "French", 3: "English")

Resultados

	precision	recall	f1-score	support
0	1.00	0.99	0.99	270
1	0.99	1.00	0.99	184
2	1.00	0.99	1.00	247
3	0.98	1.00	0.99	224
accuracy			0.99	925
macro avg	0.99	0.99	0.99	925
weighted avg	0.99	0.99	0.99	925

(Nota: Las clases son 0: "Spanish", 1: "Italian", 2: "French", 3: "English")

Concluciones

- La detección de idioma es un componente fundamental en modelos de NLP.
- La comprensión de procesos como la vectorización de texto es crucial para el desarrollo de modelos más complejos.
- Los modelos desarrollados son efectivos y pueden ser la base para aplicaciones más avanzadas.