Tarea 3. Diseño de experimentos para análisis de texto

Helena Patricia Carrillo Soto
Facultad de Ciencias Físico-Matemáticas, UANL

Maestría Ciencia de Datos: Procesamiento y Clasificación de Datos

Resumen

Se analizaron varias canciones de diez artistas diferentes. De dicho análisis se obtienen y comparan estadísticas descriptivas de texto así como se intenta clasificar cada una de las canciones con su respectivo artista.

Keywords: Artistas; Análisis de texto; Canciones

1. Introducción

Análisis a realizar En este artículo se realizará un análisis de diversas canciones por diez diferentes artistas. Se busca encontrar si es posible, y cual es la mejor forma de hacerlo, entrenar un modelo para clasificar las canciones de cada artista basándose en sus letras.

Canciones elegidas En su mayoría se eligieron canciones de artistas pop o de hip-hop los listados son músicos ampliamente conocidos. Se tienen artistas que son conocidos por escribir su música y artistas que abiertamente se apoyan de terceros para la escritura de sus canciones.

2. Descripción de los datos

Los artistas a analizar son los siguientes:

- ***Billie Eilish***. Conocida por escribir sus canciones junto con su hermano y productor Finneas O'Connell.
- ***Charlie Puth***. Artista altamente involucrado en la escritura y produccion de sus canciones.
- ***Coldplay***. La mayoría de sus canciones son escritas por su vocalista Chris Martin.
- ***Drake***. Es de conocimiento público que utiliza varios escritores para la creación de sus canciones.
- ***Ed Sheeran***. Reconocido por escribir no solo sus canciones sino también por haber escrito múltiples canciones para otros artistas.
- ***Eminem***. Conocido por escribir la mayoría de sus canciones y ha escrito para otros artistas.
- ***Justin Bieber***. Usa diferentes escritores y productores para la creación de la mayoría de sus canciones.
- ***Khalid***. Conocido por escribir y producir la mayoría de sus canciones.
- ***Rihanna***. Utiliza diferentes compositires en la escritura de sus canciones.
- ***Taylor Swift***. Mundialmente conocida por escribir la gran mayoría sus canciones.

3. Metodología

Para realizar el análisis de las canciones primero se obtuvieron varias canciones de los artistas de la base de datos en Kaggle que se puede encontrara en la siguiente liga: https://www.kaggle.com/datasets/deepshah16/song-lyrics-dataset .

Posteriormente se separaron en frases o versos y se obtuvieron algunos estadísticos como la cantidad de palabras, la cantidad de caracteres y la densidad de palabras promedio de cada artista. Después se quitaron las palabras vacías y se lematizaron los textos para obtener las palabras más características de cada verso. Para esto se utiliza la librería `nltk`.

Luego se analizaron usando una regresión logística con tres tipos de vectorizaciones y dos diferentes divisiones de entrenamiento y prueba.

Los tipos de vectorizadores que se usaron son los siguientes:

- `CountVectorizer` usando el analizador de palabras. Convierte un texto en una matriz de conteo de características.
- `TfidfVectorizer` usando ek analizador de palabras. Convierte un texto en una matriz de características TF-IDF el cuál se utiliza para medir que términos son más relevantes.
- `TfidfVectorizer` usando el analizador de caracteres.

Así mismo se usaron dos divisiones de entrenamiento y prueba: 80-20 y 70-30.

4. Resultados

Análisis Estadístico Los primeros estadísticos que se obtienen son el número de palabras, caracteres y la densidad de palabras promedio. Los resultados se pueden observar en el cuadro junto con los resultados de cada uno de los artistas 1.

De los resultados de dicho cuadro observamos que los raperos Drake y Eminem tienen considerablemente más palabras promedio que el resto de los artistas. Esto tiene sentido por el tipo de música que producen. Sin embargo la densidad de palabras no varía significativamente entre cada uno de ellos.

Cuadro 1: Número de palabras, caracteres y densidad de palabras promedio por artista.

Artist	word_count	char_count	word_density
Billie Eilish	31.298	152.480	4.960
Charlie Puth	34.486	166.524	4.888
Coldplay	26.788	130.653	5.0375
Drake	64.903	317.682	5.009
Ed Sheeran	44.631	214.032	4.802
Eminem	76.860	385.649	5.153
Justin Bieber	38.872	187.159	4.967
Khalid	45.503	224.854	5.071
Rihanna	40.336	192.679	4.998
Taylor Swift	33.402	164.103	4.984

Podemos observar los resultados gráficamente en las figuras 1 , 2 y 3

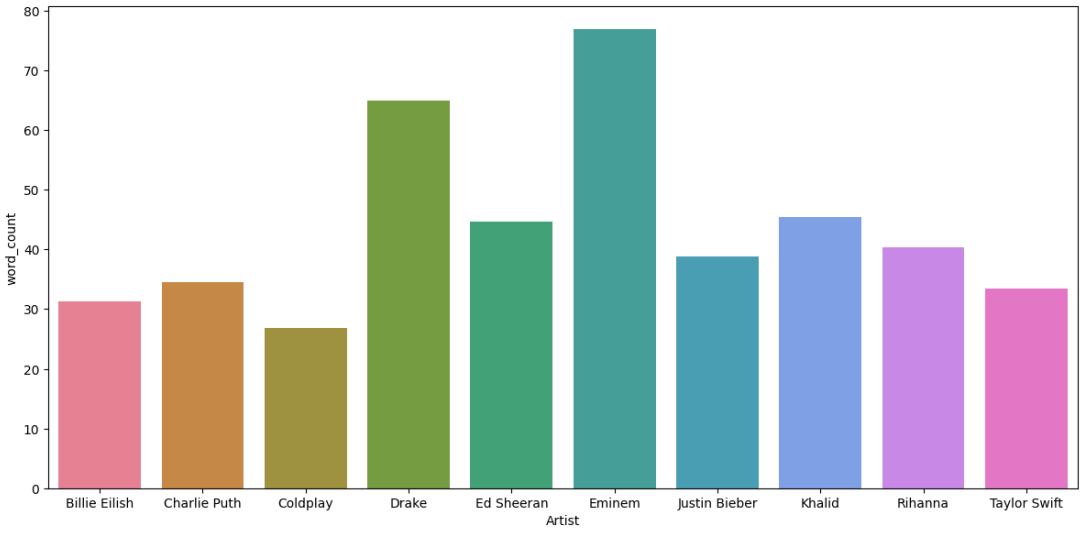

Figura 1: Palabras promedio por artista

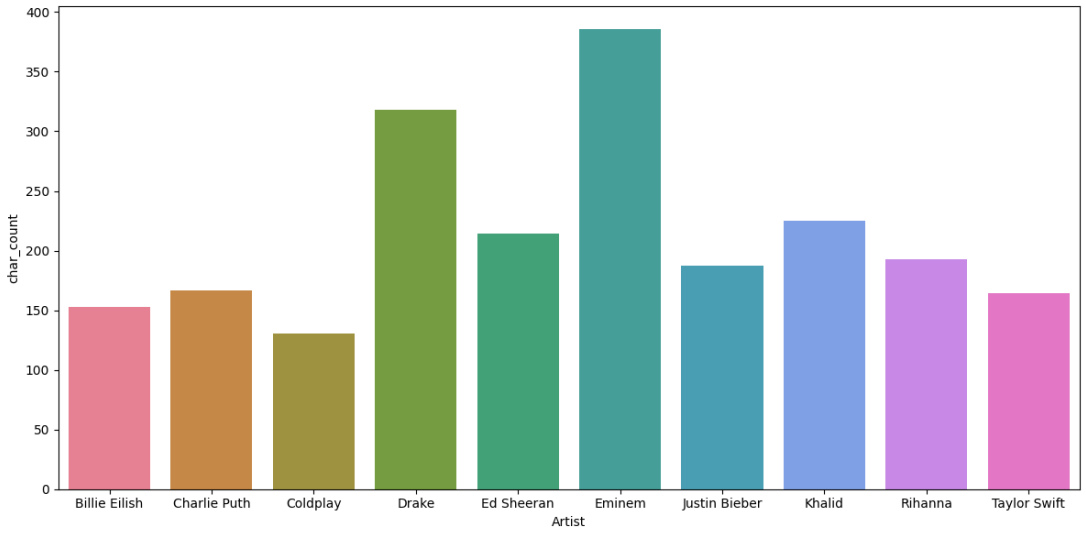

Figura 2: Caracteres promedio por artista

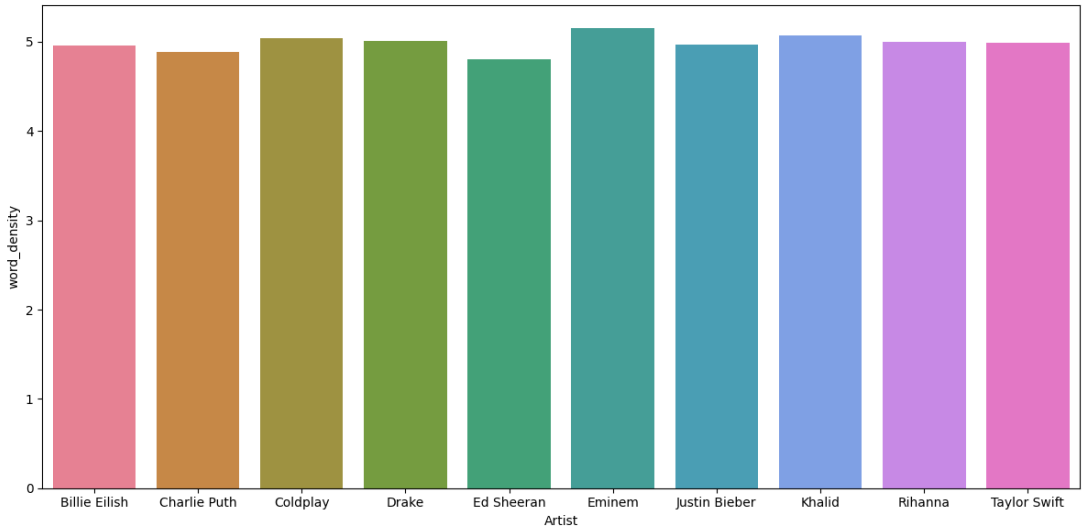

Figura 3: Densidad de palabras por artista

Posteriormente se realizó un experimento usando una regresión logística para clasificar las letras de las canciones en sus respectivos artistas.

Se usaron los tres tipos de vectorizaciones y dos tipos de división de entrenamiento y prueba antes mencionados para evaluar cuál es el que tiene el mejor desempeño evaluando el porcentaje de letras correctamente clasificadas. Los resultados se presentan en la tabla 2.

Cuadro 2: Resultados de modelos probados

Vectorización	Correctas	Tiempo	Tamaño test
CountVectorizer	0.753	10.085	0.2
CountVectorizer	0.747	9.983	0.3
CountVectorizer	0.745	9.319	0.2
Count Vectorizer	0.741	7.792	0.3
Tfidf Palabra	0.706	8.490	0.3
Tfidf Palabra	0.703	10.724	0.2
Tfidf Palabra	0.702	8.804	0.2
Tfidf Palabra	0.700	9.745	0.3
Tfidf Caracteres	0.301	1.520	0.3
Tfidf Caracteres	0.296	1.441	0.2
Tfidf Caracteres	0.294	1.132	0.2
Tfidf Caracteres	0.286	0.957	0.3

En la tabla podemos ver que el mejor método de vectorización es el de `CountVectorizer` y ya que no hay una gran diferencia entre usar una división 80-20 o 70-30 de conjuntos de entrenamiento y prueba se decide utilizar una división 80-20. Como resultado se elige un modelo de regresión logística con division 80-20 de conjunto de entrenamiento y prueba, usando la vectorización provista por la función `CountVectorizer` en la librería `sklearn`.

Al aplicar el modelo elegido se obtiene la matriz de confusión que se puede observar en la imagen 4.

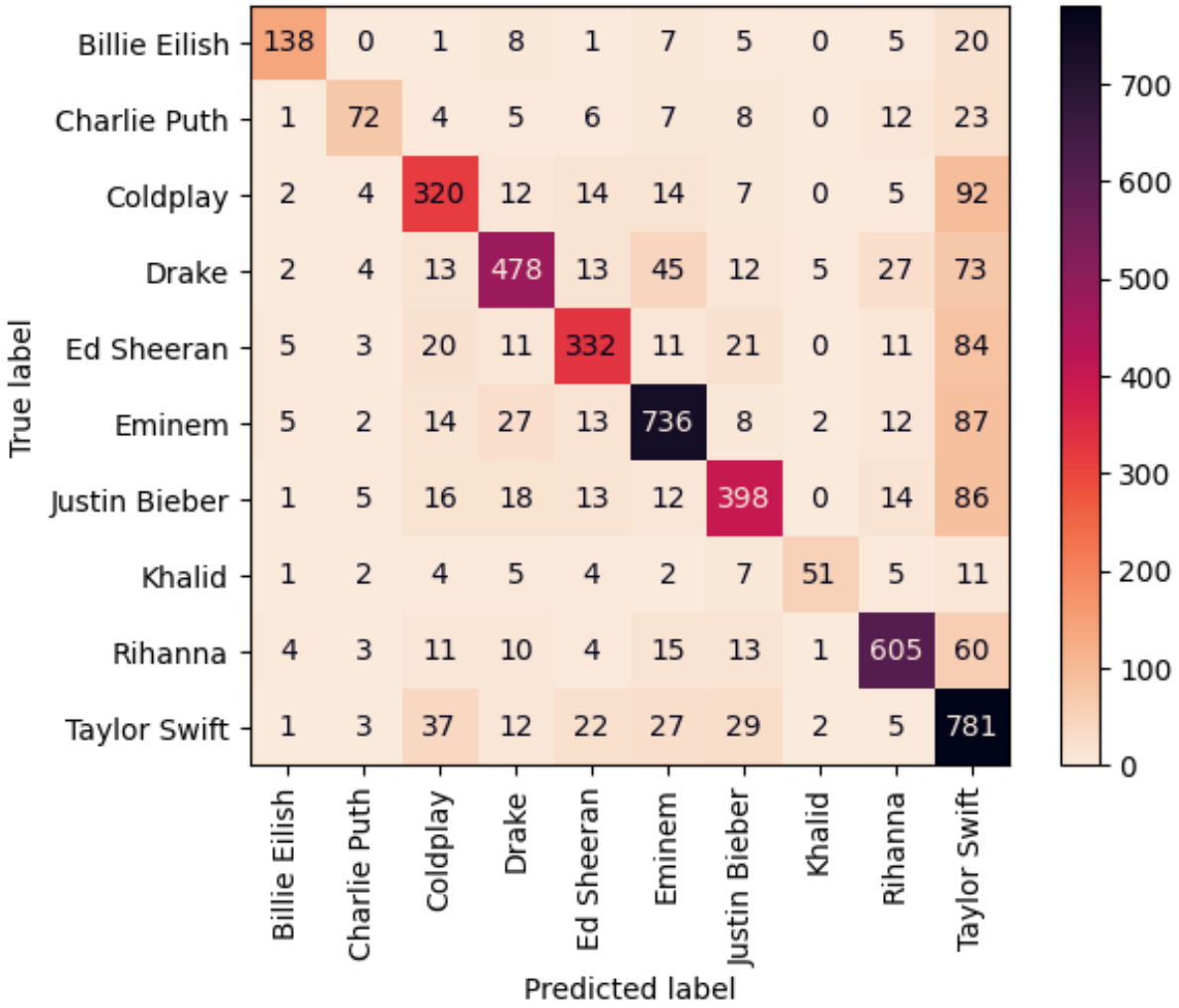

Figura 4: Matriz de confusión

5. Conclusiones

De la matriz de confusión podemos ver que el modelo efectivamente predice el artista original la mayoría de las veces. Además parece ser que la mayoría de la veces que se predice mal a un artista la predicción es asignada a Taylor Swift mientras que las canciones de Taylor Swift son mal asignadas de forma bastante uniforme entre cada artista. Puede que se deba a la similitud de temas a tratar en las canciones.