CLASIFICACIÓN DE GÉNEROS MUSICALES COLOMBIANOS UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL

Sergio Steven Lopez Martinez[1]

Abstract— **This project seeks to publicize the richness of popular Colombian music, through the study of 180 musical pieces belonging to six of the best-known Colombian musical genres: Bambuco, Carranga, Cumbia, Joropo, Pasillo, Vallenato. The study of these genres is considered through the analysis of the content of songs by means of natural language processing techniques and machine learning. We present the origin of the data set, the pre-processing tasks and the models proposed and analyzed for the subsequent evaluation of the classification results.**

Abstract— **Este proyecto busca dar a conocer la riqueza de la música popular colombiana, a través del estudio de 180 piezas musicales pertenecientes a seis de los géneros musicales colombianos más conocidos: Bambuco, Carranga, Cumbia, Joropo, Pasillo, Vallenato. El estudio de estos géneros se considera a través del análisis del contenido de las canciones mediante técnicas de procesamiento de lenguaje natural y aprendizaje automático. Presentamos el origen del conjunto de datos, las tareas de preprocesamiento y los modelos propuestos y analizados para la evaluación posterior de los resultados de la clasificación.**

I. INTRODUCCIÓN

En la música colombiana podemos encontrar una clara expresión de la cultura del país, logrando mediante su difusión compartir algunos rasgos característicos de nuestra sociedad. En este trabajo se quiere por medio del estudio de las letras de estos géneros musicales resaltar el valor cultural que poseen y motivar a muchos futuros artistas a componer nuevos temas pertenecientes a estos. Esto por medio de una descripción de características de las composiciones de las letras de los diferentes géneros, que permitirán identificar patrones o estilos de composición específicos.

II. DESCRIPCIÓN DEL PROBLEMA

La clasificación automática de música es un área importante y que ha ido aumentado su investigación en los últimos años en lo que se conoce como Musical Information Retrieval(MIR) o recuperación de información musical, principalmente mediante el estudio y aplicación de diferentes modelos de aprendizaje automático que combinan datos de audio, líricos y culturales. Sin embargo muchas de estas investigaciones tratan de abordar el problema de manera global sobre los géneros mas amplios y que están presentes en cualquier parte del mundo, siendo esto muy útil pero que a la vez no resaltan géneros específicos de una región, por lo cual el enfocar este análisis sobre los géneros musicales tradicionales de Colombia permite especificar a mas detalle que tipos de características presentan los mismos que los hacen diferentes entre si y posiblemente a otros géneros a nivel mundial. Las técnicas de aprendizaje automático han demostrado ser bastante exitosas en la extracción de tendencias y patrones de conjuntos de datos, en este caso particular el problema es netamente de procesamiento de lenguaje natural. Por eso el objetivo principal de este proyecto es mediante la utilización de técnicas de procesamiento de lenguaje natural, explorar la riqueza gramatical de estos géneros y la información valiosa que puede tener para su uso en clasificación.

[1]S. lopez. Estudiante de Ingeniería de Sistemas y computación, Universidad Nacional de Colombia sslopezm@unal.edu.co

III. SOLUCIÓN PROPUESTA

Múltiples autores han realizado pruebas sobre únicamente conjuntos de datos de letras de canciones, obteniendo buenos resultados. Por mi parte deseo abordar algunas de las características explicadas por estos autores para ser utilizadas dentro de un modelo que esta basado en las estadísticas del texto, sin embargo en el desarrollo de este trabajo pude identificar que estas estadísticas no me daban información detallada sobre los temas tratados en las canciones, ni el léxico o vocabulario característico con el que suelen estar las canciones. Teniendo en cuenta esto decidí realizar una separación en dos modelos diferentes, con el fin de poder comparar la importancia entre la información obtenida por medio de la estructura del texto y por medio de los temas y palabras utilizadas comúnmente dentro de cada genero:

III-A. Modelo Tf-idf:

Un modelo común para este tipo de problemas involucra el modelo de bolsa de palabras o Bag of Words, haciendo que cada termino que aparezca en cualquier documento de la colección sea tratado como una característica; para llenar estos vectores se realiza un conteo para calcular la frecuencia de aparición de los términos en los documentos. Para este caso como nuestro problema es de clasificación optamos por el esquema Tf-idf, o frecuencia de termino por frecuencia de documento inversa. Este modelo me permite realizar una ponderación de las palabras importantes, donde podríamos afirmar que son aquellas locamente comunes y globalmente raras, que me permiten discriminar entre documentos. Para calcular esta ponderación seguimos la siguiente formula:

$$tf \times idf(t,d) = tf(t,d) \cdot ln(N/df(t))$$

Donde d es un documento, t un token, N el numero de documentos, la frecuencia tf(t,d) es el numero de veces que aparece t en el documento d y la frecuencia de documento df(d) es la cantidad de documentos en la que aparece el termino t.

III-B. Modelo estadístico de texto:

Los documentos de texto puedes describirse por medio de estadísticas basadas en frecuencias de caracteres y palabras. Características como la complejidad de un texto pueden estar relacionadas con medidas de un vocabulario extenso o reducido, el repetir constantemente una palabra o frase; pueden ser indicadores de ciertos tipos de texto y en nuestro caso de géneros musicales. Por esta razón se plantea un modelo netamente estadístico para analizar si es posible identificar rasgos característicos de los géneros en estos aspectos.

En este modelo las features o características a entrenar están basadas en lo mencionado anteriormente y en el trabajo realizado por Mayer[1]; son las siguientes:

Feature	Metodo
PalabrasPorLinea	# Palabras / # Lineas
PalabrasUnicasPorLinea	# PalabrasUnicas / # Lineas
RazonPalabrasUnicas	# PalabrasUnicas / # Palabras
CaracteresPorPalabra	# Caracteres / # Palabras
Sentimiento	Analisis brindado por TextBlob

Decidí agregar una característica mas, no ligada a la estadística del texto como tal pero que puede aportar información del mismo, se trata de una análisis de sentimiento realizado por medio del uso de TextBlob

IV. SECCIÓN EXPERIMENTAL

IV-A. DISEÑO

El apartado experimental es realizado por medio de Colaboratory, un entorno gratuito de Jupyter Notebook que no requiere configuración y que se ejecuta completamente en la nube. Se utilizo las librerías de nltk para los procesos de lenguaje natural y sklearn para la implementacion de los modelos de aprendizaje de maquina.

IV-B. DATOS

Se utilizaron 180 archivos de texto recopilados de manera manual, estos datos requirieron ciertos pasos de preprocesamiento con el fin de tener un formato de tokens limpios que serían útiles para realizar nuestra tarea de procesamiento de lenguaje natural y posterior clasificación. Esto se realizó en varios pasos:

IV-B.1. Eliminación de ruido: Se menciona la eliminación de ruido debido a que es importante tenerla en cuenta si se desean manejar datos extraídos de paginas web ya que se debe eliminar encabezados de archivos de texto, etiquetas HTML, XML, metadatos o extraer información valiosa en el manejo de otros formatos como JSON,sin embargo al ser nuestro conjunto de datos armado manualmente no se requirió de este paso.

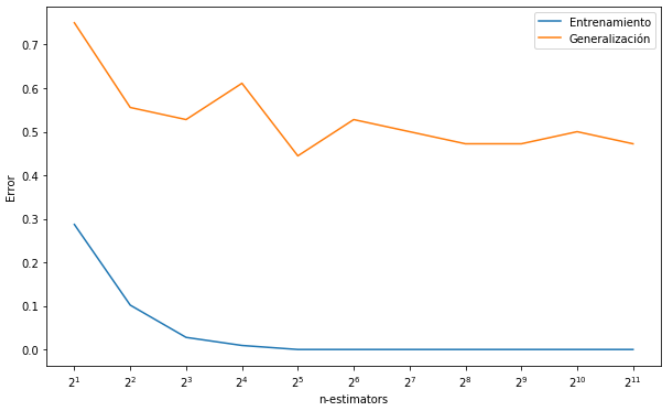
IV-B.2. Tokenizacion: Cambien conocida como análisis léxico, la tokenizacion es el proceso que divide cadenas de texto en piezas mas pequeñas o token. Este paso fue realizado por medio de la librería de nltk.

IV-B.3. Normalizacion: La normalizacion es el proceso relacionado a convertir todo el texto a un modelo determinado, esto incluye convertir todas las letras en minúscula o mayúscula, eliminar la puntuación, convertir los números en palabras equivalentes, todo con el fin de realizar el procesamiento de manera uniforme. En nuestro caso especifico la normalizacion incluirá la lematizacion y el steeming de las palabras.

IV-C. MODELO Y CONFIGURACIONES DEL MODELO

IV-D. Clasificacion Supervisada

IV-D.1. RandomForest: El primer método que utilizamos, RandomForest, es un algoritmo de aprendizaje muy certero que consiste en una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos, construyendo una larga colección de árboles no correlacionados para luego promediarlos. Es necesario hacer un analisis de complejidad de error para distintos valores de estimadores, esto con el fin de encontrar el mejor valor que ayude a reducir el error de generalizacion del algoritmo. Aqui una demostracion de este analisis y el valor elegido del cual saldran los resultados de su clasificacion posteriormente. Se eligió un valor optimo de n=5 debido a que fue el punto de menor error de generalización visto en el gráfico.



IV-D.2. Naive Bayes Multinomial: Es uno de los modelo probabilistas mas simples y utilizados por sus resultados tan buenos como otros métodos mas sofisticados. Se basa en la aplicación de la Regla de Bayes para predecir probabilidades condicionales de que los documento pertenezca a una clase P(A/B) a partir de la probabilidad de los documentos dada la clase P(B/A) y la probabilidad a priori.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

Este modelo no observa oraciones en su totalidad sino que supone que cada palabra es independiente una de a otra, eliminando problemas de probabilidades de 0 en frases que no se encuentran en el conjunto de entrenamiento. Finalmente calcula las probabilidades y compara cual es la mas alta.

IV-D.3. LinearSVC: El objetivo de un SVC lineal (Support Vector Classifier) es adecuarse a los datos de entrenamiento proporcionados, devolviendo un hiperplano ideal que divide o categoriza los datos de forma optima. Su diferencia con los SVM(Support Vectro Machine) esta en el kernel utilizado, es por esto que se especifica en el nombre que en el caso de los SCV su kernel es lineal. La característica fundamental de las vectores de soporte es que buscan el hiperplano que tenga la máxima distancia con los puntos que estén más cerca de él mismo. De esta forma, los puntos del vector que son etiquetados con una categoría estarán a un lado del hiperplano y los casos que se encuentren en la otra categoría estarán al otro lado.

IV-D.4. Regresion Logistica: La Regresion Logistica es una tecnica de aprendizaje automatico que para clasificacion; podriamos definirla como una red neuronal con exactamente 1 neurona. La regresion logistica tiene dos partes, una combinacion lineal a la izquierda de la neurona y una aplicacion de la funcion logistica(sigmoidal) la derecha de la neurona. Es decir combina todas nuestras entradas y luego aplica a dicho resultado la funcion sigmoidal. La cual esta dada por la siguiente ecuacion:

$$\sigma = \frac{1}{1 + e^{-z}}$$

IV-D.5. Red Neuronal: El perceptron multicapa (Multilayer Perceptron, MLP en inglés) es un tipo especial de red neuronal en el cual se apilan varias capas de perceptrones. Para nuestra red neuronal, se utiliza keras, que nos permite definir el modelo base y agregarle capas a medida que se requiera, en nuestro caso nuestra red está construida con 4 capas de 256, 128, 64 y 6 neuronas respectivamente. Teniendo claramente esta ultima capa una funcion de activacion Softmax.

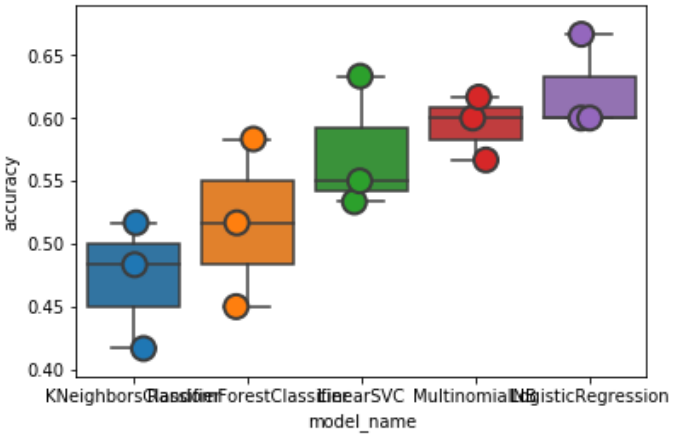
IV-E. Clasificacion no Supervisada

IV-E.1. Clustering: El objetivo del Clustering es agrupar objetos clases de objetos similares, es una tarea no supervisada, pues no sabemos a priori como clasificar nuestros objetos, por lo cual al algoritmo solo se le pasara las features de los documentos, en este caso todas las features de los pesos tf-idf y de acuerdo al analisis de estas, se agruparan con otras similares. Se utilizara un algoritmo de clustering de k-medias, es un algoritmo particional que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano

V. RESULTADOS

V-A. Modelo Tf-idf:

Comparacion acuraccy promedio de diferentes modelos:



Esta grafica nos indica la accuracy o exactitud promedio del modelo para nuestro conjunto de datos de caracteristicas Tf-idf, siendo:

```
model_name
KNeighborsClassifier      0.472222
LinearSVC                 0.572222
LogisticRegression        0.622222
MultinomialNB             0.594444
RandomForestClassifier    0.516667
Name: accuracy, dtype: float64
```

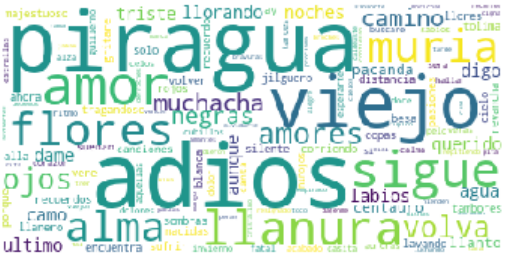
Teniendo en cuenta los resultados dados por esta evaluación de modelos, se opto por evaluar nuestro conjunto de datos con el modelo de Regresion Logistica, obteniendo los siguientes resultados:

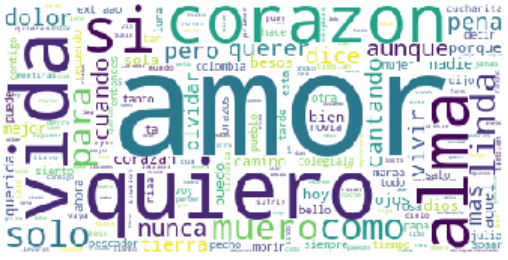
```
Accuracy: 0.5370370370370371
Error: 0.4629629629629629

Etiquetas: ['bambuco', 'carranga', 'cumbia', 'joropo', 'pasillo', 'vallenato']

Precision: [0.4        0.625      1.         0.66666667 0.75       0.41176471]
Recall:    [0.66666667 0.55555556 0.44444444 0.44444444 0.33333333 0.77777778]
F_1 score: [0.5        0.58823529 0.61538462 0.53333333 0.46153846 0.53846154]
```

Matriz de confusión

	bambuco	carranga	cumbia	joropo	pasillo	vallenato
bambuco	6	1	0	0	0	2
carranga	1	5	0	0	1	2
cumbia	2	0	4	1	0	2
joropo	2	0	0	4	0	3
pasillo	3	1	0	1	3	1
vallenato	1	1	0	0	0	7

Luego se realizo un borrado de las etiquetas de nuestro modelo para poder entrenar un clasificador no supervisado, el cual representamos en forma de nube de palabras nuestros 6 clusters:





V-B. Modelo estadístico de texto:

Clasificador RandomForest

```
Accuracy: 0.3055555555555556
Error: 0.6944444444444444

Etiquetas: ['bambuco', 'carranga', 'cumbia', 'joropo', 'pasillo', 'vallenato']

Precision: [0.         0.33333333 0.         0.375      0.5        0.55555556]
Recall:    [0.         0.33333333 0.         0.5        0.16666667 0.83333333]
F_1 score: [0.         0.33333333 0.         0.42857143 0.25       0.66666667]
```

Matriz de confusión

	bambuco	carranga	cumbia	joropo	pasillo	vallenato
bambuco	0	1	3	1	1	0
carranga	1	2	1	1	0	1
cumbia	1	1	0	2	0	2
joropo	0	2	0	3	0	1
pasillo	4	0	1	0	1	0
vallenato	0	0	0	1	0	5

Clasificador LinearSVC

```
Accuracy: 0.3888888888888889
Error: 0.6111111111111112

Etiquetas: ['bambuco', 'carranga', 'cumbia', 'joropo', 'pasillo', 'vallenato']

Precision: [1.         0.27272727 0.         0.28571429 0.66666667 0.36363636]
Recall:    [0.16666667 0.5        0.         0.33333333 0.66666667 0.66666667]
F_1 score: [0.28571429 0.35294118 0.         0.30769231 0.66666667 0.47058824]
```

Matriz de confusión

	bambuco	carranga	cumbia	joropo	pasillo	vallenato
bambuco	1	3	0	1	1	0
carranga	0	3	0	0	1	2
cumbia	0	2	0	1	0	3
joropo	0	2	0	2	0	2
pasillo	0	1	0	1	4	0
vallenato	0	0	0	2	0	4

Clasificador RedNeuronal

```
Accuracy: 0.4166666666666667
Error: 0.5833333333333333

Etiquetas: ['bambuco', 'carranga', 'cumbia', 'joropo', 'pasillo', 'vallenato']

Precision: [0.         0.44444444 0.         0.44444444 0.42857143 0.44444444]
Recall:    [0.         0.66666667 0.         0.66666667 0.5        0.66666667]
F_1 score: [0.         0.53333333 0.         0.53333333 0.46153846 0.53333333]
```

Matriz de confusión

	bambuco	carranga	cumbia	joropo	pasillo	vallenato
bambuco	0	0	1	1	3	1
carranga	0	4	0	0	1	1
cumbia	0	3	0	1	0	2
joropo	0	1	0	4	0	1
pasillo	1	1	0	1	3	0
vallenato	0	0	0	2	0	4

Clasificador Multinomial Naive Bayes

```
Accuracy: 0.3611111111111111
Error: 0.6388888888888888

Etiquetas: ['bambuco', 'carranga', 'cumbia', 'joropo', 'pasillo', 'vallenato']

Precision: [1.         0.25       0.25       0.14285714 0.57142857 0.375     ]
Recall:    [0.33333333 0.33333333 0.16666667 0.16666667 0.66666667 0.5       ]
F_1 score: [0.5        0.28571429 0.2        0.15384615 0.61538462 0.42857143]
```

Matriz de confusión

	bambuco	carranga	cumbia	joropo	pasillo	vallenato
bambuco	2	0	3	1	0	0
carranga	0	2	0	0	3	1
cumbia	0	2	1	1	0	2
joropo	0	3	0	1	0	2
pasillo	0	1	0	1	4	0
vallenato	0	0	0	3	0	3

Clasificador RegresionLogistica

```
Accuracy: 0.3333333333333333
Error: 0.6666666666666667

Etiquetas: ['bambuco', 'carranga', 'cumbia', 'joropo', 'pasillo', 'vallenato']

Precision: [0.33333333 0.44444444 0.         0.22222222 0.66666667 0.33333333]
Recall:    [0.16666667 0.66666667 0.         0.33333333 0.33333333 0.5       ]
F_1 score: [0.22222222 0.53333333 0.         0.26666667 0.44444444 0.4       ]
```

Matriz de confusión

	bambuco	carranga	cumbia	joropo	pasillo	vallenato
bambuco	1	0	3	2	0	0
carranga	0	4	0	0	1	1
cumbia	0	2	0	1	0	3
joropo	0	2	0	2	0	2
pasillo	2	1	0	1	2	0
vallenato	0	0	0	3	0	3

VI. DISCUSIÓN DE RESULTADOS

Los resultados de los metodos de apendizaje maquinal utilizados reflejan un bajo porcentaje de exactitud, siendo el mejor de los casos la clasificación por medio de regresión logística para el modelo Tf-idf que obtuvo un accuracy del 53,7 %. Mientras que para los métodos utilizados para la evaluacion del modelo de estadísticas de texto el accuracy obtenido oscila entre el 30 % y el 42 %.

Esta diferencia puede atribuirse a la cantidad de características presentadas en cada modelo, donde el modelo de Tf-idf con la cantidad de 180 canciones y calculando el tf-idf solo para palabras que aparecieran mas de 4 veces en la colección obtuvo 1156 características para entrenar, sin embargo considero que la diferencia de accuracy no es tan significativa si tenemos en cuenta que el modelo de estadísticas de texto solo esta evaluando 5 características. Por tanto un buen ejercicio seria realizar una evaluación de un modelo combinado que permitiera al modelo de estadísticas conocer mas características de los documentos. Sin duda alguna hay que considerar que uno de los factores importantes de estos géneros esta en sus ritmos e instrumentos que los componen, por lo cual también es importante tener características dadas por la parte del audio de las piezas estudiadas. Algo a destacar es los resultados obtenidos específicamente para la clase de Vallenato, siendo esta la clase mas consistente en todos los modelos, reflejando que las palabras utilizadas en el genero y la forma en que están constituidas sus letras es muy única de su genero, contando historias largas y en la mayoría de los casos como dedicatorias de amor o desamor. Siendo estas identificaciones lo buscado en el análisis de estos textos.

Si nos remontamos a los métodos de clasificación no supervisada se obtuvo una separación en grupos que permitió una identificación coherente de diferentes temas, por lo cual considero que el análisis y clasificación de letras por técnicas de procesamiento de lenguaje natural y agrupamiento puede utilizarse para la detección de tópicos o temas en canciones de todo tipo.

VII. TRABAJO A FUTURO

Una tarea fundamental para poder mejorar los resultados obtenidos es sin duda la ampliación del conjunto de datos. Pues este trabajo solo fue una prueba de la viabilidad de la aplicacion de estas tecnicas a los generos musicales representativos de Colombia, ademas de una dificil obtencion de los documentos debido a la gran influencia de otros generos musicales actuales, por lo cual la produccion de estos generos culturales no es muy amplia y su informacion en la web con respecto a sus letras es aun menor. A pesar de tratarse y abordarse como un problema de procesamiento de lenguaje natural, seria interesante poder combinar este enfoque con un estudio y análisis de audio, pues esto brindaría una cantidad considerable de características que permita una mejor identificación de los géneros, pues podemos observar con los resultados que en muchos casos las características de estadísticas de textos y la gran cantidad de palabras representativas de las regiones hacen que no haya una clara correlación entre las letras y sus géneros.

REFERENCIAS

1 Mayer, Rudolf, Robert Neumayer and Andreas Rauber. “Rhyme and Style Features for Musical Genre Classification by Song Lyrics.” Proc. of International Society for Music Information Retrieval. ISMIR (2008).

2 Hu, X., Downie, J.S., Ehmann. ”Lyric text mining in music mood classification.İn: Proc. of International Society for Music Information Retrieval. ISMIR (2009)

3 Laurier, C., Grivolla, J., Herrera, P.: Multimodal music mood classification using audio and lyrics. In: Proc. of International Conference on Machine Learning and Applications. (2008)

4 van Zaanen, M., Kanters, P.: Automatic mood classification using tf-idf based on lyrics. In: Proc. of International Society for Music Information Retrieval. (2010)

5 Buzic, Dalibor Dobša, Jasminka. (2018). Lyrics Classification Using Naive Bayes.