Descubrimiento de conocimiento de Datos Complejos

Autores: Marcos Sanchez, Ignacio Franco y Carlos Menéndez

1. Introducción

La música popular en español constituye un espejo privilegiado de los cambios sociales, culturales y lingüísticos de una sociedad. En particular, durante las últimas décadas se ha observado una transformación notable en los estilos y contenidos de las letras musicales más populares. Esta transformación se manifiesta en una progresiva adopción de un lenguaje más coloquial y con una presencia creciente de expresiones explícitas.

En nuestro proyecto la principal dificultad reside en la operacionalización del concepto de «vulgaridad». Tres factores contribuyen significativamente a esta complejidad: primero, se trata de una categoría fundamentalmente subjetiva y altamente dependiente del contexto temporal, puesto que las expresiones consideradas vulgares en la década de 1940 no coinciden necesariamente con aquellas percibidas como tales en la actualidad; segundo, el vocabulario experimenta una evolución constante, con la aparición continua de nuevas expresiones que requieren reinterpretación; y tercero, los datos disponibles presentan un desequilibrio significativo entre períodos, concentrando una proporción abrumadoramente mayor de composiciones en años recientes, lo que introduce un sesgo potencial en cualquier comparación temporal.

Se plantea la siguiente hipótesis: desde la década de 1940 hasta 2025, las letras de las canciones más populares y buscadas en español han experimentado un incremento progresivo en la presencia de expresiones vulgares y explícitas.

El objetivo central del proyecto es responder de manera objetiva y fundamentada la siguiente pregunta: ¿han aumentado cuantitativamente las expresiones vulgares y explícitas en las letras de canciones populares en español conforme han transcurrido las décadas?

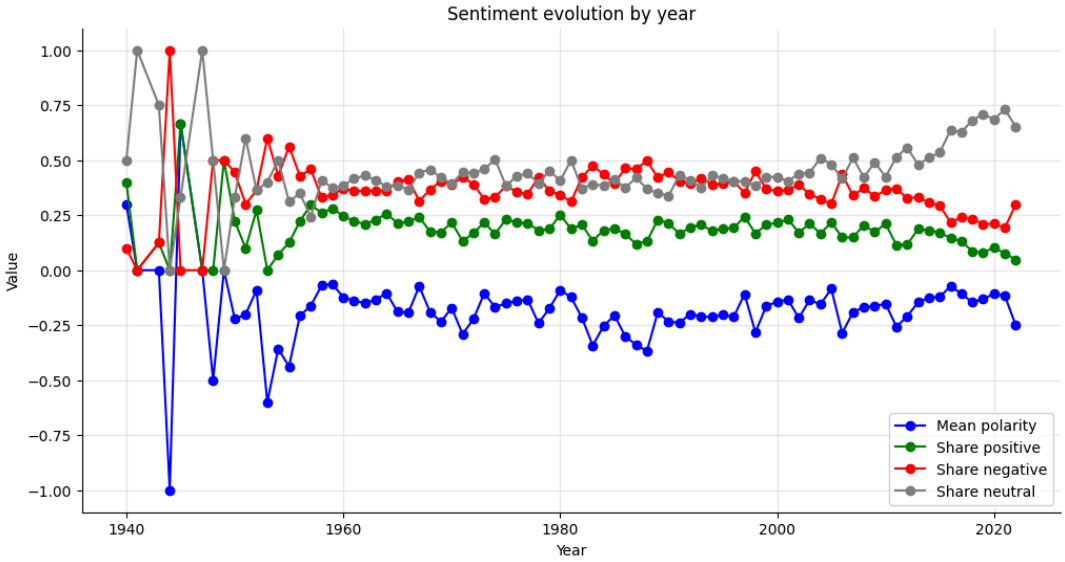
2. Metodología

Se utilizó un dataset extenso de letras en español con metadatos de año, popularidad y reproducciones. Tras depuración exhaustiva de registros temporales erróneos, se delimitó el análisis al período 1940-actualidad.

Para mitigar el sesgo de concentración de canciones recientes, se implementó muestreo estratificado, seleccionando aproximadamente 1,690 composiciones por década con prioridad en popularidad y distribución de géneros comparable entre períodos.

El texto se limpió eliminando marcadores estructurales, tokens onomatopéyicos y stopwords, aplicando posteriormente lematización con spaCy para reducir variabilidad morfológica del español.

Se entrenó un modelo LdaSeqModel con 7 tópicos para capturar la evolución temporal de temas principales en letras musicales. Los tópicos se interpretaron manualmente mediante sus términos más representativos y visualizaciones en WordClouds, facilitando la identificación de patrones semánticos a lo largo de las décadas.



También implementamos un análisis de sentimiento mediante la herramienta pysentimiento, que clasifica el contenido textual en categorías de polaridad: positiva (POS), neutra (NEU) y negativa (NEG). Estos valores se agregaron temporalmente para obtener métricas resumen como la polaridad media (mean_polarity), permitiendo examinar si la evolución hacia un lenguaje más vulgar se acompaña de cambios en la tonalidad emocional de las composiciones.

La validación temporal de la hipótesis se abordó mediante un enfoque multiescalar. En primer lugar, se calcularon estadísticos descriptivos del indicador de vulgaridad para cada año, aplicando posteriormente una suavización mediante media móvil para reducir fluctuaciones de corto plazo y evidenciar tendencias subyacentes. Complementariamente, se realizaron comparaciones entre décadas, evaluando métricas tales como la proporción de composiciones que superan un umbral predefinido de vulgaridad.

Para caracterizar la evolución temporal con mayor precisión, se implementaron modelos de series temporales, específicamente modelos SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables) e, hipotéticamente, alternativas como Holt-Winters, con particiones formales train/test para validación robusta.

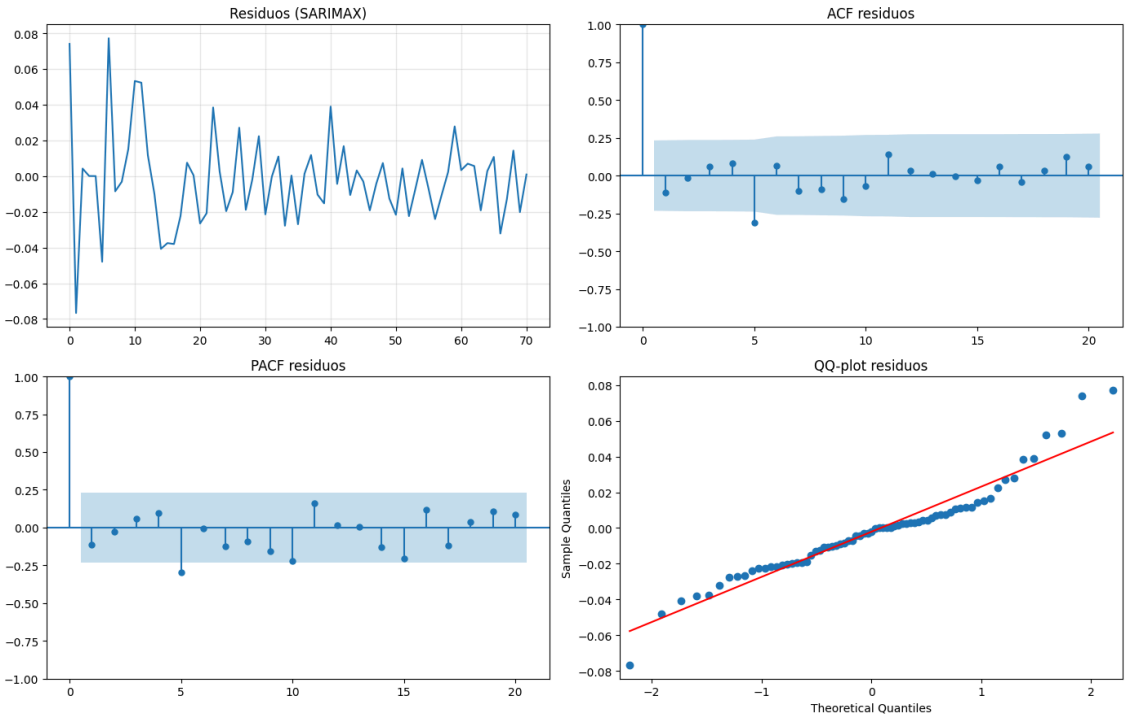
También realizamos un análisis de volatilidad mediante el cálculo de retornos del indicador vulgar y la estimación de modelos GARCH (Generalized AutoRegressive Conditional Heteroskedasticity), permitiendo la detección de períodos de cambio acelerado o "ráfagas" de transformación lingüística.

3. Justificación Metodológica del Modelo SARIMAX

Se modeló la evolución temporal de la vulgaridad mediante una serie temporal anual, construida agregando el peso medio del tópico vulgar por año. Esta transformación convierte la caracterización cualitativa del tono de las letras en una señal cuantitativa continua, directamente comparable a lo largo del período de estudio.

Hemos optado por el modelo SARIMAX dado que permite capturar simultáneamente: (i) la no estacionariedad y tendencias mediante diferenciación; (ii) dependencias temporales mediante términos autorregresivos (AR) y de media móvil (MA); y (iii) la posibilidad de incorporar variables exógenas adicionales (otros tópicos, sentimiento, etc.). En el caso base, el objetivo fue validar que el patrón temporal resultaba capturable mediante un modelo estándar y establecer una referencia de error para comparativas posteriores.

El análisis incluyó: el orden del modelo (p,d,q)(P,D,Q,s), que especifica la estructura de memoria y diferenciación; criterios de información AIC y BIC para comparación y selección de configuraciones; coeficientes estimados de componentes AR, MA y estacionales con sus respectivos p-valores, verificando la significancia de cada término; diagnóstico de residuos mediante funciones de autocorrelación (ACF/PACF) y prueba de Ljung–Box, confirmando la ausencia de autocorrelación residual; y finalmente, métricas de desempeño fuera de muestra (MAE/RMSE) junto con visualizaciones de predicciones versus valores observados, demostrando la capacidad de generalización del modelo.



La especificación del modelo SARIMAX se desarrolló mediante un enfoque parsimonioso, partiendo de una configuración base que fue refinada iterativamente conforme a criterios técnicos específicos. En particular, se incorporó diferenciación (d=1) para estabilizar la media de la serie, se añadieron componentes autorregresivos y de media móvil de orden moderado (evitando órdenes elevados que indujeran sobreajuste, dada la limitación en el tamaño de la serie), y se incluyó un término estacional ligero para capturar patrones cíclicos de corto plazo sin forzar estructuras estacionales innecesariamente complejas.

La selección final se basó en la evaluación sistemática de múltiples configuraciones cercanas, priorizando aquella que balanceara óptimamente tres criterios: (i) valores bajos de AIC/BIC; (ii) coeficientes estimados con interpretación económica y significancia estadística; y (iii) diagnósticos de residuos satisfactorios.

Para garantizar que las métricas de error reflejasen desempeño genuino y evitar sesgos de autoevaluación, se implementó una partición temporal explícita train/test, entrenando el modelo con observaciones históricas y evaluándolo en un bloque final reservado. A partir de las predicciones en el conjunto de prueba se calcularon métricas de error: MAE (error absoluto medio), interpretable como desviación típica en unidades del indicador vulgar, y RMSE (raíz del error cuadrático medio), que penaliza proporcionalmente los errores de mayor magnitud.

Interpretación del RMSE y sus Implicaciones

Un RMSE aproximado de 0.11 unidades indica que, en promedio, el modelo se desvía aproximadamente 0.11 unidades del indicador de vulgaridad. Este resultado debe contextualizarse considerando tres factores críticos: primero, la serie contiene relativamente pocas observaciones anuales, limitando inherentemente la capacidad del modelo para aprender patrones complejos; segundo, existe un quiebre estructural marcado a partir del año 2000, intensificándose tras 2015, que viola la suposición de dinámica suave que los modelos SARIMAX típicamente presuponen; tercero, por tanto, aunque SARIMAX mejora respecto a métodos de suavizado simple, los errores deben interpretarse como una métrica de seguimiento temporal más que como predicción fiable de saltos estructurales abruptos.

En contraste, el método Holt-Winters captura adecuadamente tendencias y suavizado, pero al no modelar explícitamente autocorrelación y shocks exógenos, presenta una capacidad inferior para reaccionar ante cambios de pendiente, resultando en un error de prueba superior al del SARIMAX.

Consideraciones Finales sobre la Interpretación

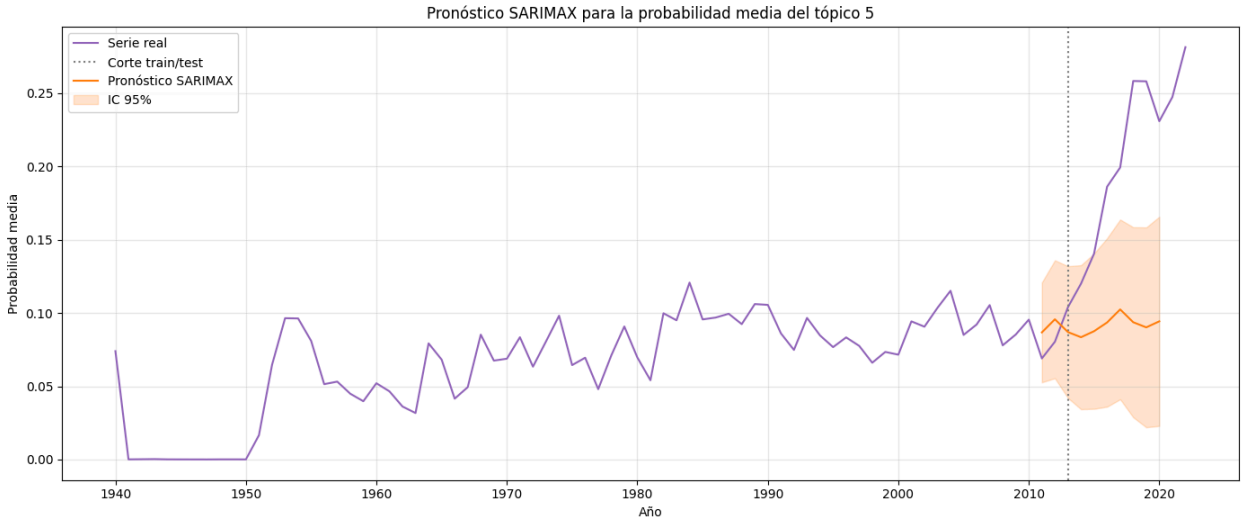
Las métricas de RMSE y MAE no se presentan como validación absoluta del modelo, sino como medidas objetivas de ajuste fuera de muestra en una serie caracterizada por cambio estructural. Precisamente esta limitación metodológica refuerza la conclusión central del análisis: el aumento de vulgaridad observado en períodos recientes

constituye un quiebre cualitativo respecto a la dinámica histórica previa, no una simple prolongación de tendencias previas.

4. Resultados

Hemos ajustado el modelo SARIMAX con especificación (1,1,1)(0,1,1,5), utilizando un período de entrenamiento hasta 2012 y reservando años posteriores para validación. El desempeño en el conjunto de prueba muestra un RMSE de 0.10974 y MAE de 0.09565, indicadores de error moderado consistentes con la naturaleza de la serie. Los criterios de información resultaron AIC = -292.56 y BIC = -284.32.

Respecto a la significancia de los coeficientes, el componente MA(1) resultó estadísticamente significativo ($p \approx 0.023$), mientras que el término AR(1) no alcanzó significancia ($p \approx 0.644$). Este patrón es típico en series temporales breves: el modelo captura la tendencia subyacente con relativamente pocos términos, mientras que algunos coeficientes permanecen poco identificados debido a la limitada información disponible para su estimación.



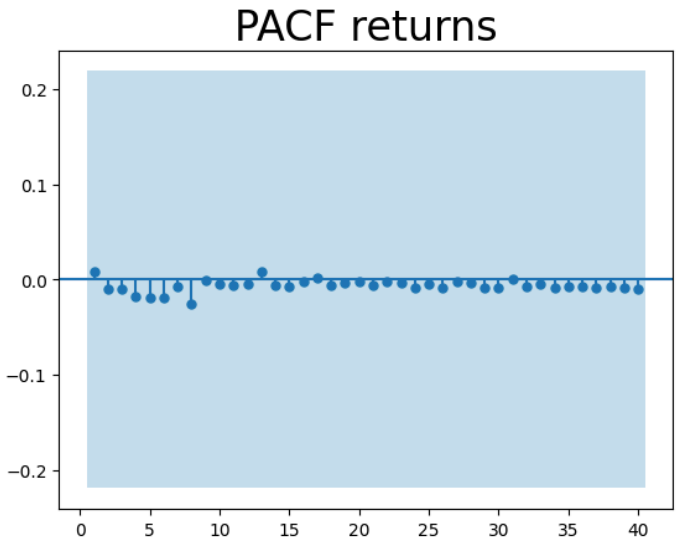
Diagnóstico de Residuos y Análisis de Volatilidad

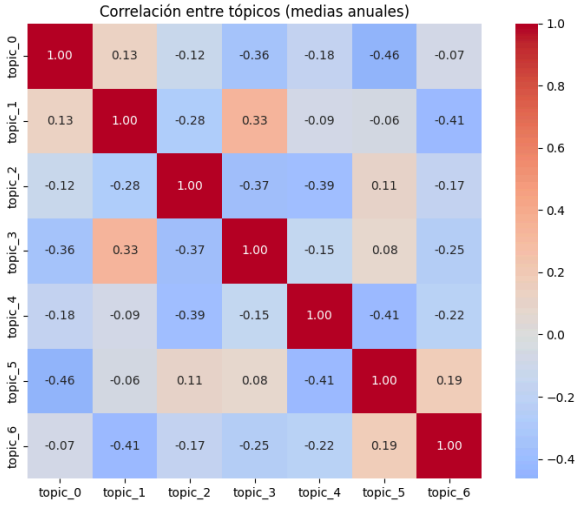
La prueba de Ljung–Box ($p \approx 0.186$) confirma ausencia de autocorrelación residual significativa. Los gráficos ACF/PACF y QQ-plot corroboran que el modelo captura adecuadamente la estructura temporal.

El modelo SARIMAX con variables exógenas (topic_0 y topic_4) produjo una mejora modesta en RMSE (0.10487) y MAE (0.09078), pero el BIC empeoró (-280.03) al penalizar la complejidad adicional, indicando que la ganancia predictiva es limitada.

El análisis GARCH(1,1) sobre retornos reveló un coeficiente de persistencia de varianza aproximadamente 0.89, evidenciando clustering de volatilidad en la serie. Se

aplicó estandarización previa para mitigar el impacto de valores extremos en períodos iniciales de baja magnitud del indicador.





5. Justificación de elección de parámetros

Componente no estacional (1,1,1).

Se seleccionó d=1 porque la serie del tópico 5 exhibe tendencia no estacionaria; esta diferenciación mejora el ajuste sin complejidad innecesaria. Los órdenes p=1 y q=1 fueron elegidos por parsimonia, logrando error de prueba razonable (RMSE ≈ 0.11) con residuos que pasan validación básica (Ljung–Box p ≈ 0.186).

Componente estacional (0,1,1,5).

Se incorporó D=1 para capturar variación de mediano plazo no controlada solo con d=1. Los órdenes P=0 y Q=1 responden a parsimonia: la serie es breve (~80 años) y órdenes elevados generan sobreajuste. Aunque el término estacional puede no ser individualmente significativo, se mantiene como componente ligero si favorece estabilidad numérica e intervalos de confianza sin deteriorar el desempeño en test.

Métrica de selección (AIC/BIC versus RMSE/MAE).

Mientras RMSE/MAE cuantifican generalización en el bloque de prueba, AIC/BIC comparan modelos penalizando complejidad adicional. En la especificación con exógenas, la mejora en RMSE contrastó con empeoramiento en BIC, justificando la conclusión de mejora "ligera pero no concluyente" como trade-off entre ajuste y parsimonia.

Selección de exógenas (topic_0 y topic_4).

Se eligieron por máxima correlación absoluta con topic_5. El resultado —mejora pequeña en error pero BIC peor— sugiere relación limitada, insuficiente para afirmar superioridad clara sin validación temporal más robusta.

6. Interpretación

Modelo SARIMAX Base.

El RMSE ≈ 0.11 y MAE ≈ 0.096 representan errores medios en unidades del indicador. Dado que el tópico 5 oscila entre ~0.08 en períodos históricos y ~0.28 en años recientes, el modelo tiende a subestimar el quiebre post-2015, siendo útil como referencia comparativa mas no como predicción absoluta. La significancia de MA(1) frente a AR(1) no significativo sugiere que la dependencia capturable es principalmente de corto plazo (shocks), mientras que los p-valores elevados del componente estacional indican débil identificación en esa dimensión. Cabe destacar que AIC/BIC solo permiten comparación entre modelos sobre los mismos datos, sin certificar bondad absoluta del ajuste.

Modelo con Exógenas.

La mejora en RMSE/MAE al incorporar topic_0 y topic_4 evidencia información compartida entre tópicos, coherente con la matriz de correlaciones. Sin embargo, el empeoramiento del BIC y advertencias de convergencia sugieren ganancia moderada que no compensa claramente la complejidad añadida. Conclusión: existe relación, pero requiere validación adicional antes de afirmar superioridad.

Análisis GARCH.

El enfoque mediante retornos² es metodológicamente correcto para evidenciar clustering de volatilidad; un α elevado implica persistencia de varianza. Una limitación crítica es que el indicador toma valores pequeños inicialmente, generando retornos extremos que dominan el ajuste. Con ~80 observaciones anuales, GARCH debe interpretarse como análisis exploratorio, no como estimación robusta.

Conclusión Integral.

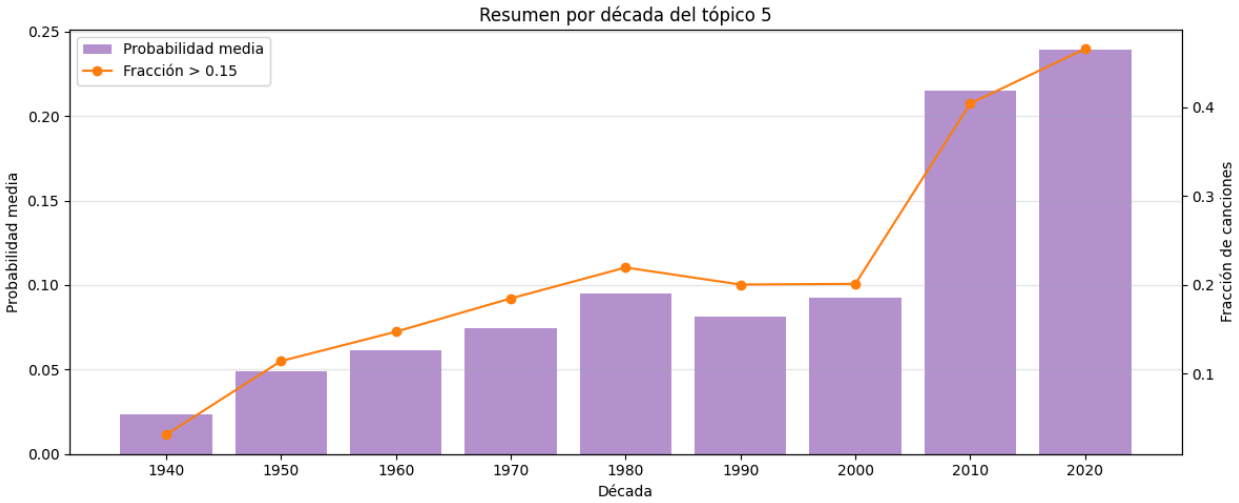
Los modelos capturan tendencia y dependencia general, pero el cambio estructural pronunciado (2000-2020) limita la capacidad predictiva fuera de muestra. Aun así, los resultados apoyan la hipótesis: el aumento reciente constituye un quiebre cualitativo, no prolongación suave de dinámicas previas.

7. Evaluación

La evolución del peso del tópico identificado como más coloquial/vulgar muestra un incremento sostenido con un aumento especialmente marcado en décadas recientes

(2000–2020). Esto es coherente con la idea de “más explícito” en la música popular contemporánea.

Al equilibrar el número de canciones por década, el patrón general se mantiene: las décadas más recientes concentran más presencia del tópico vulgar, lo que refuerza que no es solo un efecto de “hay más canciones modernas”. El SARIMAX capta la tendencia global, pero sufre para anticipar saltos rápidos (post-2015). Justamente ese fallo es consistente con un cambio estructural reciente, no con una dinámica estable sin cambios.



8. Contribución

El presente proyecto desarrolla un pipeline reproducible y transferible para el análisis temporal de letras musicales populares en español. Su arquitectura integral comprende: (i) limpieza textual rigurosa, normalización lingüística y lematización mediante herramientas estándar; (ii) agregación de datos a nivel anual y decadal para facilitar análisis longitudinales; y (iii) documentación completa de cada etapa para garantizar reproducibilidad futura.

Una contribución metodológica central es el control sistemático del sesgo temporal inherente a corpus musicales. Mediante equilibrado estratificado por décadas, se asegura una representación aproximadamente uniforme de canciones por período, evitando la falacia de concluir incrementos de vulgaridad basándose únicamente en la mayor disponibilidad de registros recientes.

Se construyó un proxy cuantitativo de vulgaridad mediante Dynamic Topic Modeling (LdaSeq), identificando un tópico específico caracterizado por vocabulario coloquial y explícito. El peso medio de este tópico por año constituye un indicador temporal

comparable, transformando un concepto cualitativo en una señal numérica susceptible de análisis formal.

La evaluación temporal se fundamentó en modelos interpretables: SARIMAX validado con métricas fuera de muestra (RMSE/MAE), criterios de información (AIC/BIC) y diagnósticos de residuos (Ljung–Box, ACF/PACF). Se evaluaron especificaciones con variables exógenas (otros tópicos), demostrando relaciones moderadas pero sin mejoras concluyentes tras considerar trade-offs de complejidad. Complementariamente, se realizó análisis exploratorio de volatilidad mediante GARCH(1,1), presentado con cautela dadas las limitaciones del tamaño muestral.

9. Conclusión general

En conjunto, el análisis sugiere que el lenguaje de las letras populares en español ha evolucionado hacia un registro **más coloquial y explícito**, con un incremento más claro en las décadas recientes. Al transformar letras en indicadores temporales (tópicos dinámicos) y estudiar su comportamiento con modelos de serie temporal, observamos un patrón compatible con la hipótesis: el componente asociado a “vulgaridad/explicitud” crece con el tiempo y muestra señales de cambio estructural en la etapa contemporánea.

No obstante, la conclusión debe entenderse como evidencia basada en un **proxy** y en un corpus concreto de canciones. Por ello, no se debe interpretar como una medida absoluta o universal de “vulgaridad” en toda la música en español.