

Newspaper Archive Sentiment

PROYECTO FINAL - RAFAEL RODRÍGUEZ

Motivación

¿Qué ideologías influyen en los periódicos que leemos?

¿Qué tan neutrales son las noticias?

¿Cómo varía el análisis de sentimientos cuando hablan del mismo partido o miembro político? ¿Y si hablan de la monarquía o de la pandemia?





Camino

03



DATA SCRAPING

Newspaper3k es una increíble biblioteca de python inspirada en requests para extraer y preparar artículos

EXPLORACIÓN DE DATOS

Explorar y entender los datos con pandas, matplotlib o seaborn

LIMPIEZA DE DATOS Y TEXT NORMALIZATION

Uso de pandas y regex para borrar filas duplicadas, columnas que no necesitamos, símbolos raros...

SENTIMENT ANALYSYS

Uso de la librería sentiment_analysis_spanish para calcular el sentimiento de los textos que necesitamos

VISUALIZACIONES Y AGGLOMERATIVE CLUSTERING

Creaciones de visualizaciones con matplotlib, seaborn, wordcloud y algoritmo de clasificación no supervisada con sklearn

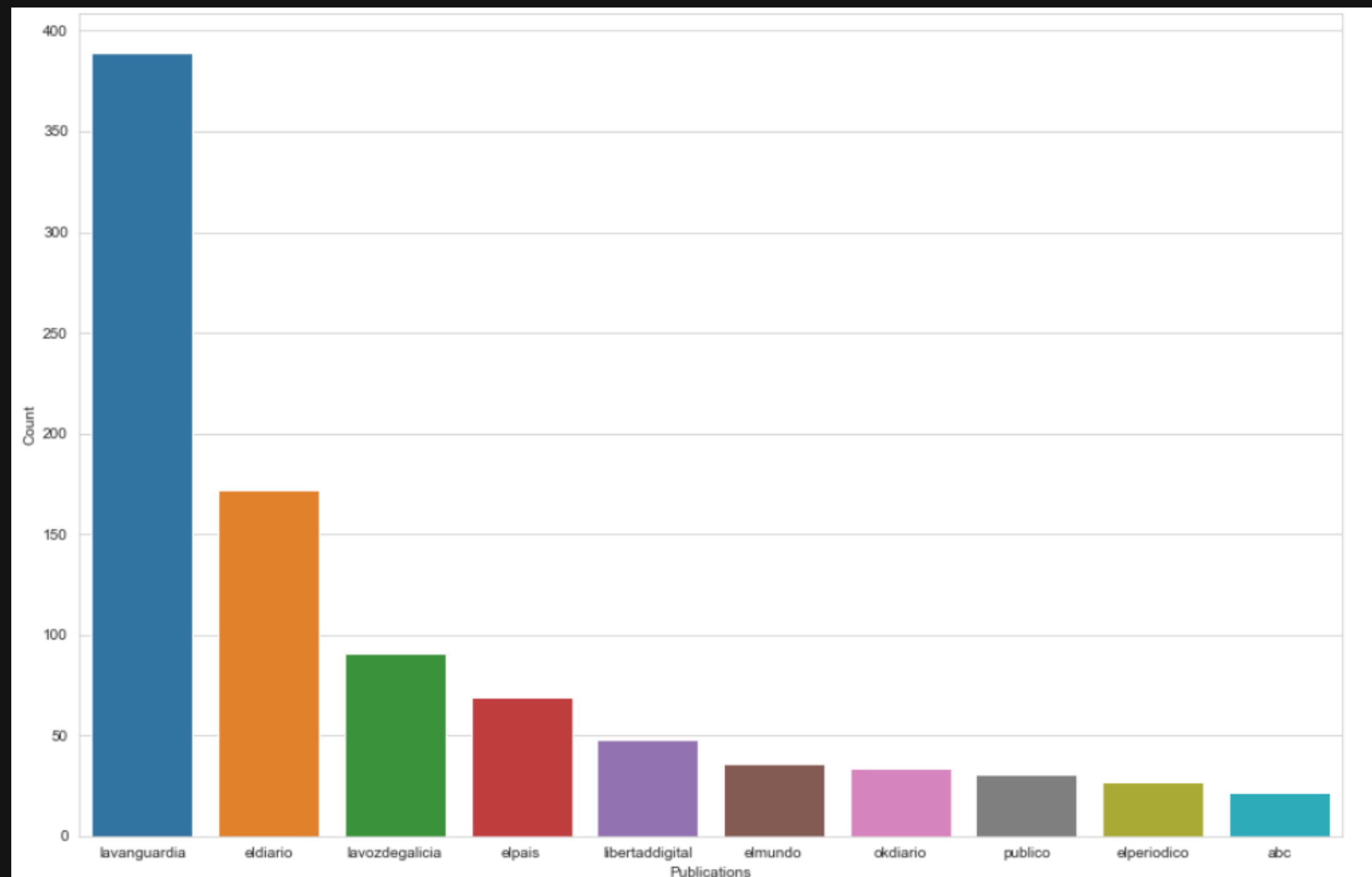


DATA

DATA: 85980

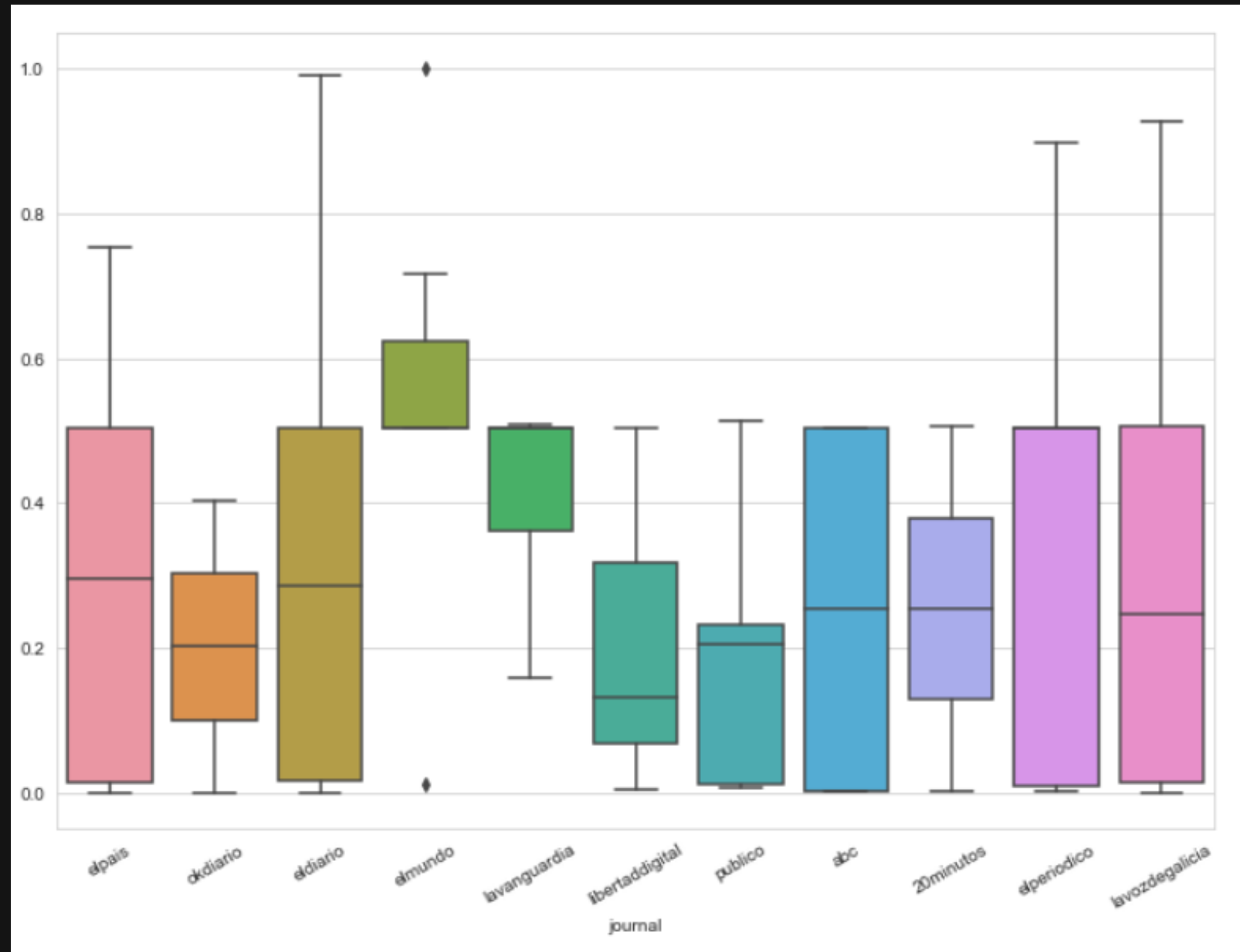
CLEAN DATA: 23032

POLITICAL DATA: 998





Casado vs Iglesias



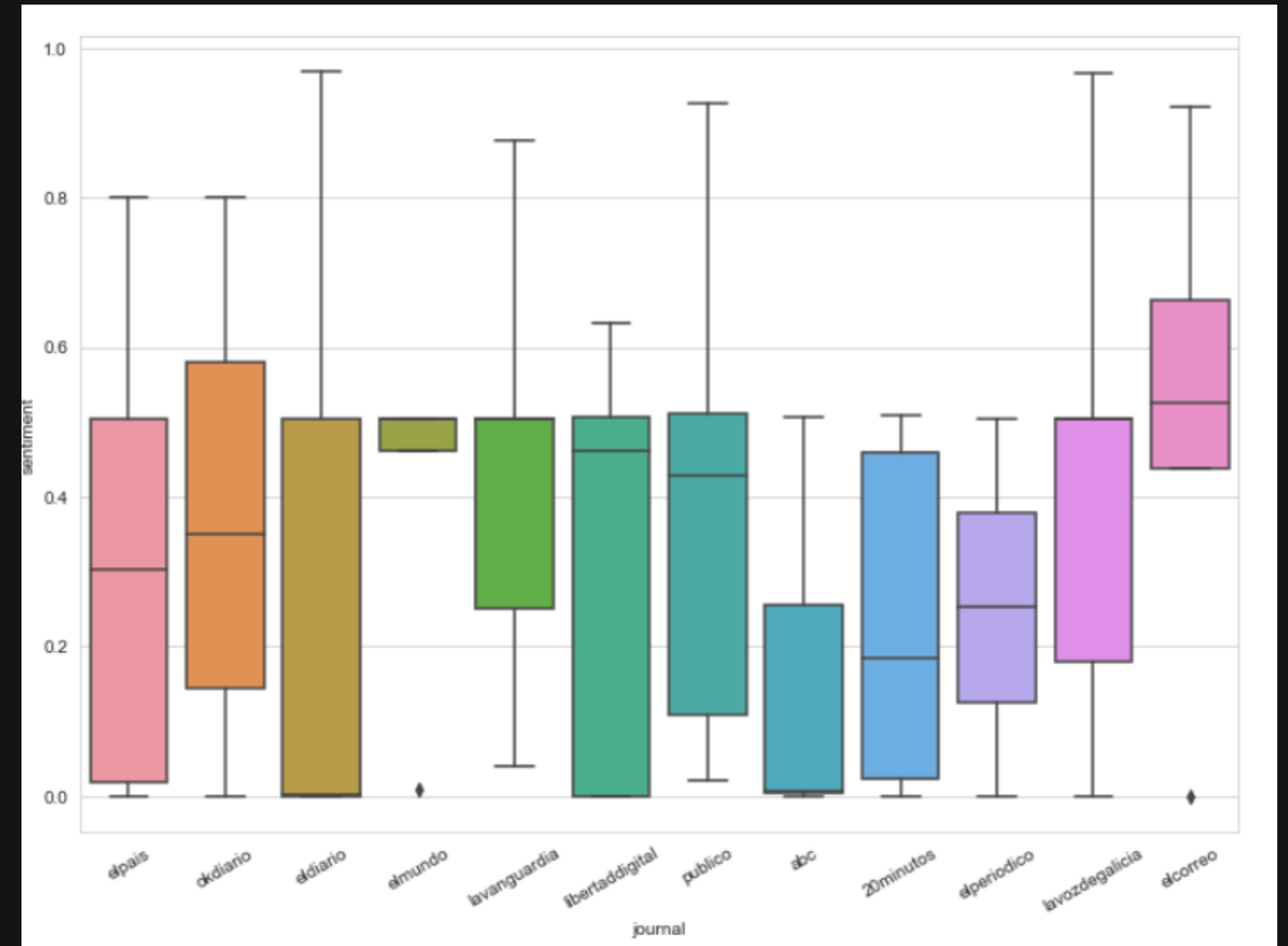
Casado

BEST_SCORE: 0.539491

JOURNAL: ELMUNDO

WORST_SCORE: 0.193913

JOURNAL: EL CORREO VASCO



Iglesias

BEST_SCORE: 0.509857

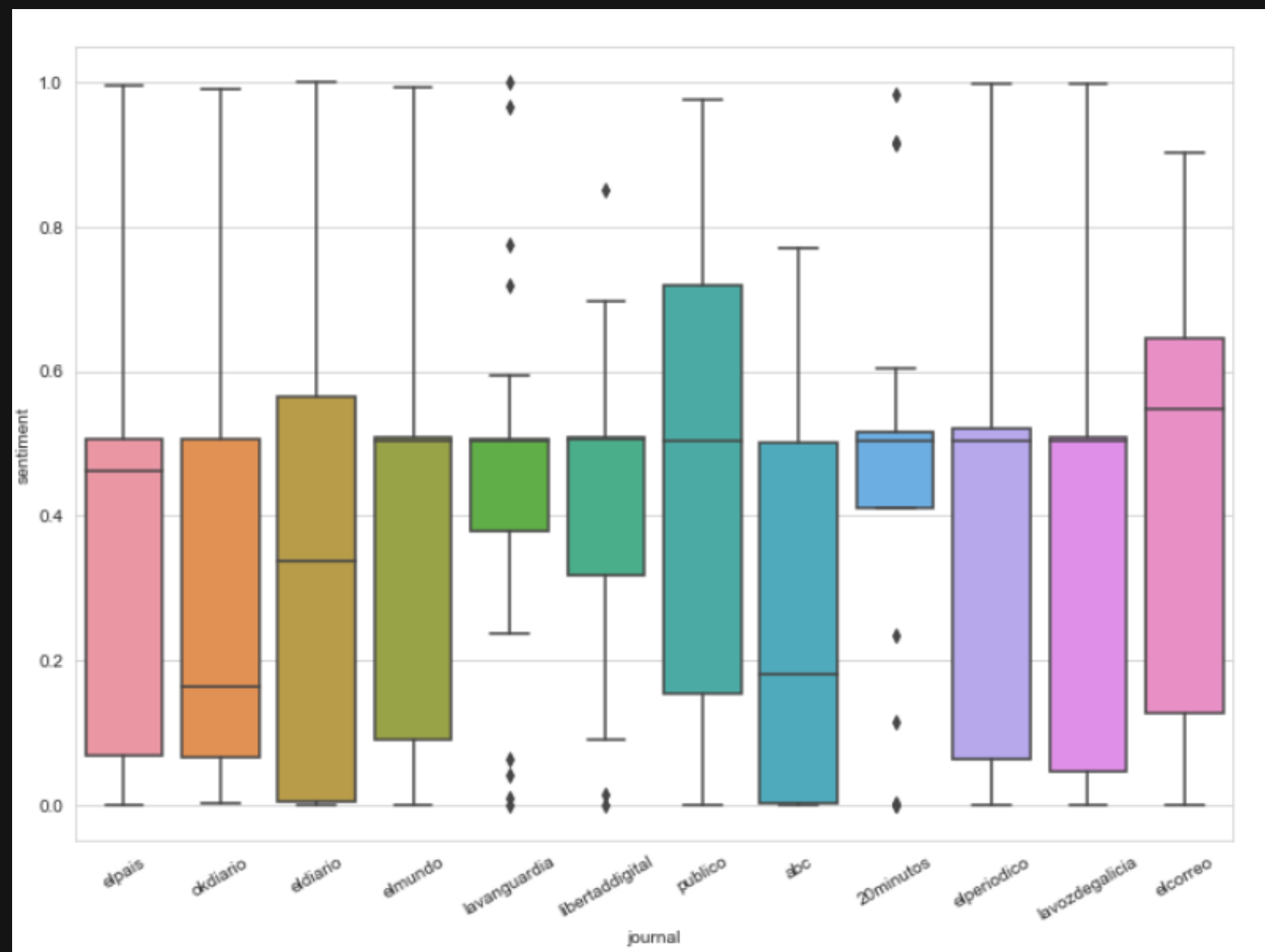
JOURNAL: EL CORREO VASCO

WORST_SCORE: 0.171052

JOURNAL: ABC



Covid-19 and Monarquía



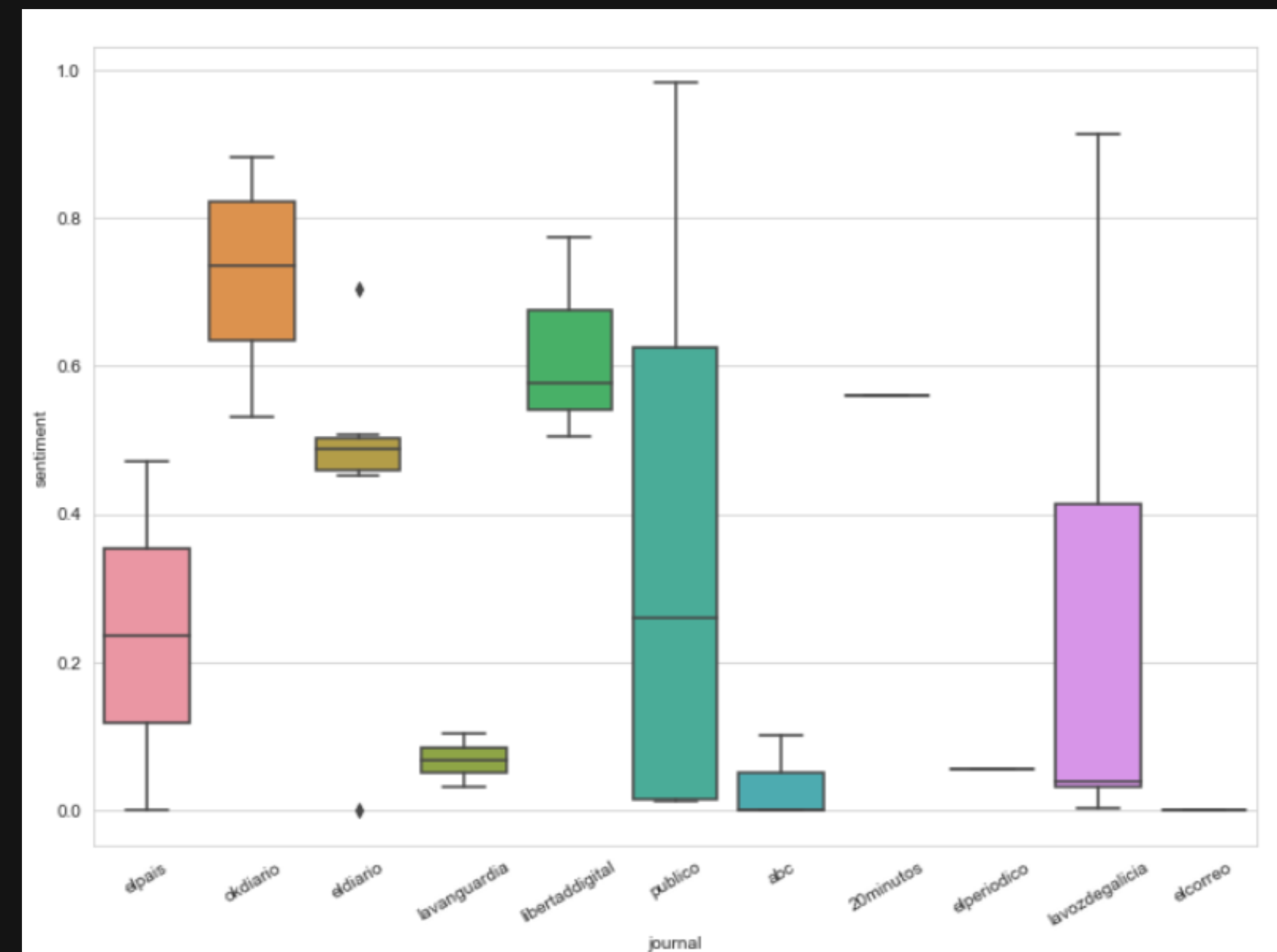
Covid

BEST_SCORE: 0.460245

JOURNAL: 20 MINUTOS

WORST_SCORE: 0.238361

JOURNAL: ABC



Monarquía

BEST_SCORE: 0.720914

JOURNAL: OK DIARIO

WORST_SCORE: 0.000192

JOURNAL: EL CORREO VASCO

Dendrogram

Euclidean distances

Journals

20minutos lavanguardia libertaddigital okdiario lavozdegalicia elmundo elperiodico elcorreo elpais publico abc eldiario

Próximos pasos

- Automatizar proceso
- Clean code
- Utilizar más fuentes de datos como rss
- Crear una aplicación donde el usuario pueda elegir la categoría y los periódicos que quiere analizar. También se podría hacer por países





MUCHAS GRACIAS
PROYECTO FINAL IRONHACK

RAFAEL RODRÍGUEZ CANCELA