Take-Home Exercise 3

2025-06-15

# Take Home Exercise 3

Link to VAST Challenge 2025, [Mini-Challenge 2](https://vast-challenge.github.io/2025/MC2.html).

# Interactive Exploration of COOTEFOO Bias: A Shiny Application Storyboard

## Background and Context

Oceanus, a small island nation, has historically relied on a robust fishing-based economy. In recent years, however, tourism has begun rapidly expanding, significantly reshaping local economic dynamics. To guide this complex transition responsibly, the government established the **Commission on Overseeing the Economic Future of Oceanus (COOTEFOO)**. This appointed oversight board, while lacking direct legislative power, holds substantial influence and funding, entrusted to evaluate economic trends objectively and advise accordingly.

COOTEFOO’s role, however, has come under intense scrutiny. Two vocal advocacy groups have recently accused the board of deep-seated biases:

* **Fishing is Living and Heritage (FILAH)** claims COOTEFOO is overly biased towards promoting tourism, neglecting the island’s historical economic foundation in fishing.
* Conversely, **Tourism Raises OceanUs Together (TROUT)** argues COOTEFOO remains stubbornly loyal to fishing interests, effectively ignoring lucrative opportunities offered by tourism.

In response to these competing accusations, investigative journalist **Edwina Darling Moray (E.D.)** of the Haacklee Herald obtained the datasets provided by both FILAH and TROUT, containing detailed meeting minutes and travel records for COOTEFOO members. Moray also secured additional evidence beyond what either advocacy group had compiled. After initial data cleaning and knowledge-graph creation by a staff programmer, these datasets were prepared for analysis.

# Investigative Narrative: Shiny Application

The interactive Shiny visualisation tool is structured to guide Moray through four distinct yet closely connected analytical stages. Each stage builds logically upon the previous, forming a coherent investigative narrative that progressively deepens understanding and systematically evaluates accusations of bias.

## Initial EDA



The Exploratory Data Analysis (EDA) dashboard provides an overview of sentiment and activity coverage across three datasets (FILAH, TROUT, JOURNALIST) to identify potential biases and gaps in COOTEFOO’s discussions.

Firstly, sentiment distributions (violin plots) show that TROUT leans positively towards tourism and has limited coverage of fishing, while FILAH presents wider sentiment variability in fishing. JOURNALIST provides a more balanced view, sitting between the two extremes.

Secondly, the stacked bar chart highlights clear gaps: TROUT under-represents fishing sentiment, while FILAH lacks data on other industry categories, emphasising skewed coverage.

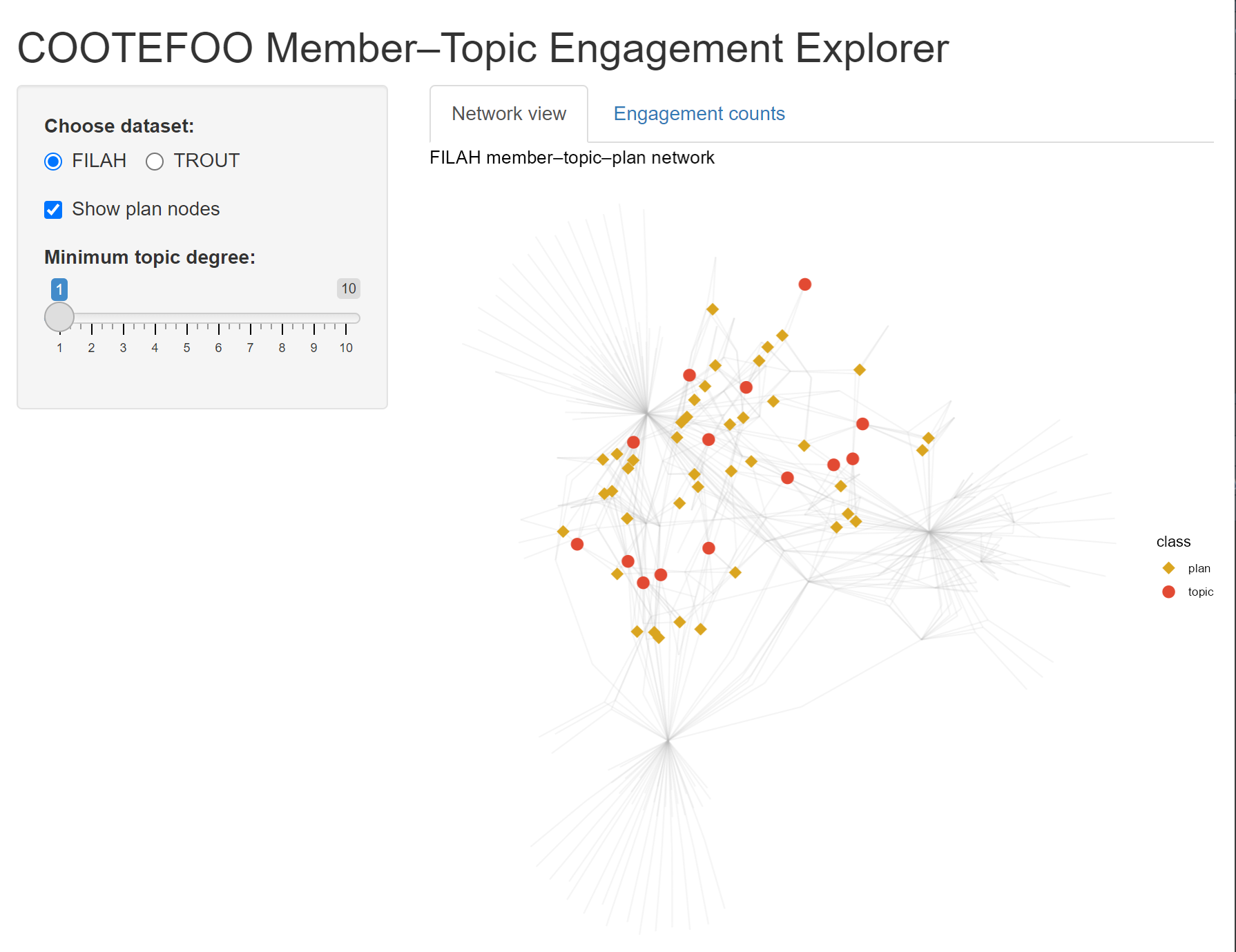
Thirdly, member activity bar charts reveal uneven representation, some members, like Simone Kat, appear frequently across all datasets, whereas others, like Seal and Teddy Goldstein, have sparse coverage in FILAH or TROUT.

Lastly, a sentiment coverage table explicitly summarises member representation across datasets, clearly illustrating individual-level gaps. Overall, the EDA reveals dataset-specific biases, justifying deeper interactive analysis in subsequent stages.

## Stage 1: Initial Claims - Examining Individual Sources (FILAH & TROUT)

Moray’s investigation begins by evaluating the datasets from TROUT and FILAH separately. Is there intrinsic evidence of bias within each group’s own records?

### Visualisation



A force-directed node-link diagram (ggraph) explicitly illustrates relationships among COOTEFOO members, discussion topics (fishing/tourism), and plans. Nodes and edges coloured by industry reveal patterns of engagement.

Interactive bar charts quantify exactly how much each member engages with fishing vs tourism topics, highlighting asymmetric attention and possible echo chambers.

**Sidebar Controls**

**Choose dataset:**

* Toggles which advocacy group’s data you’re exploring.
* FILAH shows the “Fishing Is Living and Heritage” graph, TROUT would swap in “Tourism Raises OceanUs Together.”

**Show plan nodes:**

* When checked, you see **golden-rod diamonds** representing individual policy or action plans discussed by members.
* Unchecking hides those diamonds so you focus purely on member–topic links.

**Minimum topic degree (slider = 1):**

* Filters out topics with fewer than *n* member connections.
* At “1,” all topics appear. Sliding up to, say, “3” would remove topics discussed by fewer than three members, decluttering the graph.

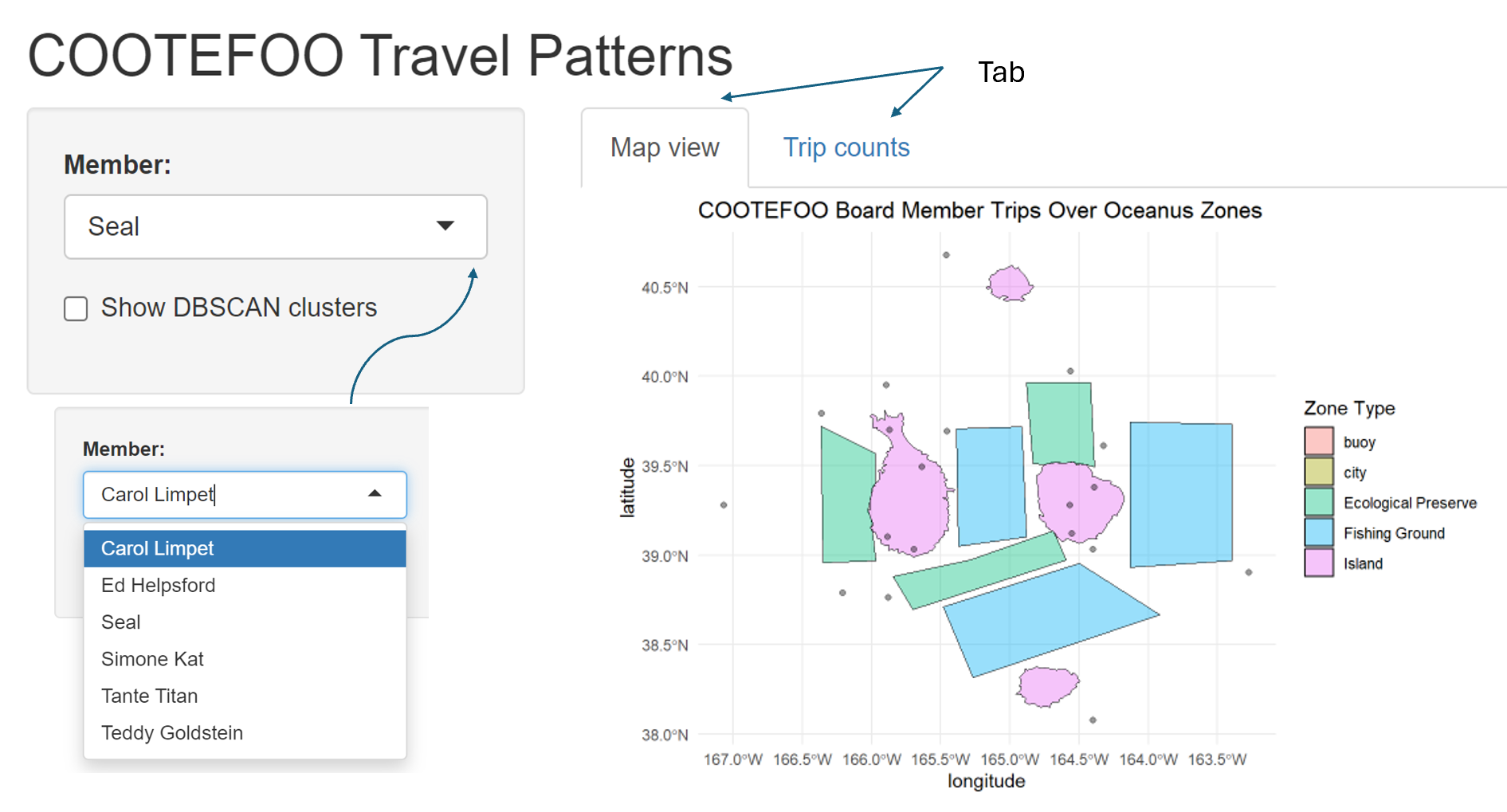
### Insights and Links to Next Stage:

Through this visualisation, Moray explicitly identifies preliminary biases. Members notably skewed toward one industry become focal points in subsequent analyses. The biases identified here set expectations for travel patterns explored in Stage 2, allowing Moray to test whether members’ stated interests match their real-world actions.

## Stage 2: Spatial Bias - Where Do COOTEFOO Members Actually Go?

Having identified initial biases in member engagement, Moray next integrates all available records (FILAH, TROUT, additional data) to examine COOTEFOO’s official travel activities. Do their physical travels reinforce or contradict biases observed in Stage 1?

### Visualisation



An interactive map (leaflet) clearly displays official COOTEFOO travel locations, explicitly marking fishing-related visits (blue markers) versus tourism-related visits (orange markers).

DBSCAN clusters explicitly identify spatial “echo chambers” or regions consistently favoured, coloured and labelled by industry type.

An accompanying interactive stacked bar chart summarises trips per member by industry, explicitly linking real-world actions to biases identified in Stage 1.

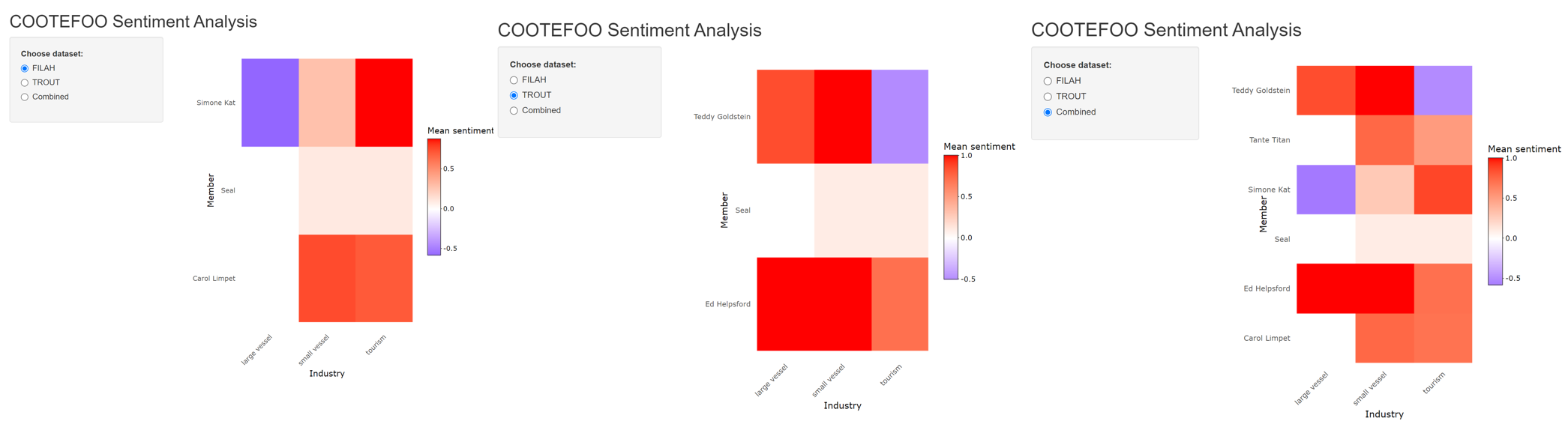
### Insights and Links to Next Stage:

Moray explicitly compares observed travel behaviour against initial engagement biases. Spatial clusters reinforce or challenge perceived biases from discussions, preparing Moray to further quantify the emotional and ideological tone behind these activities in Stage 3.

## Stage 3: Sentiment Analysis - Quantifying Emotional Tone and Bias

With clear insights into engagement and travel, Moray now examines the underlying emotional tone of COOTEFOO members’ recorded comments across the fishing and tourism contexts. Critically, Moray needs to understand how missing records from FILAH and TROUT distort or shift perceptions when compared with the combined complete dataset.

### Visualisation

Three interactive sentiment heatmaps (plotly) explicitly compare average sentiment (negative to positive) for each member across FILAH, TROUT, and the comprehensive combined dataset.

Interactive drill-down capabilities let Moray explicitly explore detailed comments underpinning sentiment scores, providing transparency and context.

Delta bars explicitly quantify changes in sentiment scores when merging additional data, indicating whether initial accusations from TROUT and FILAH become strengthened, weakened, or unchanged.

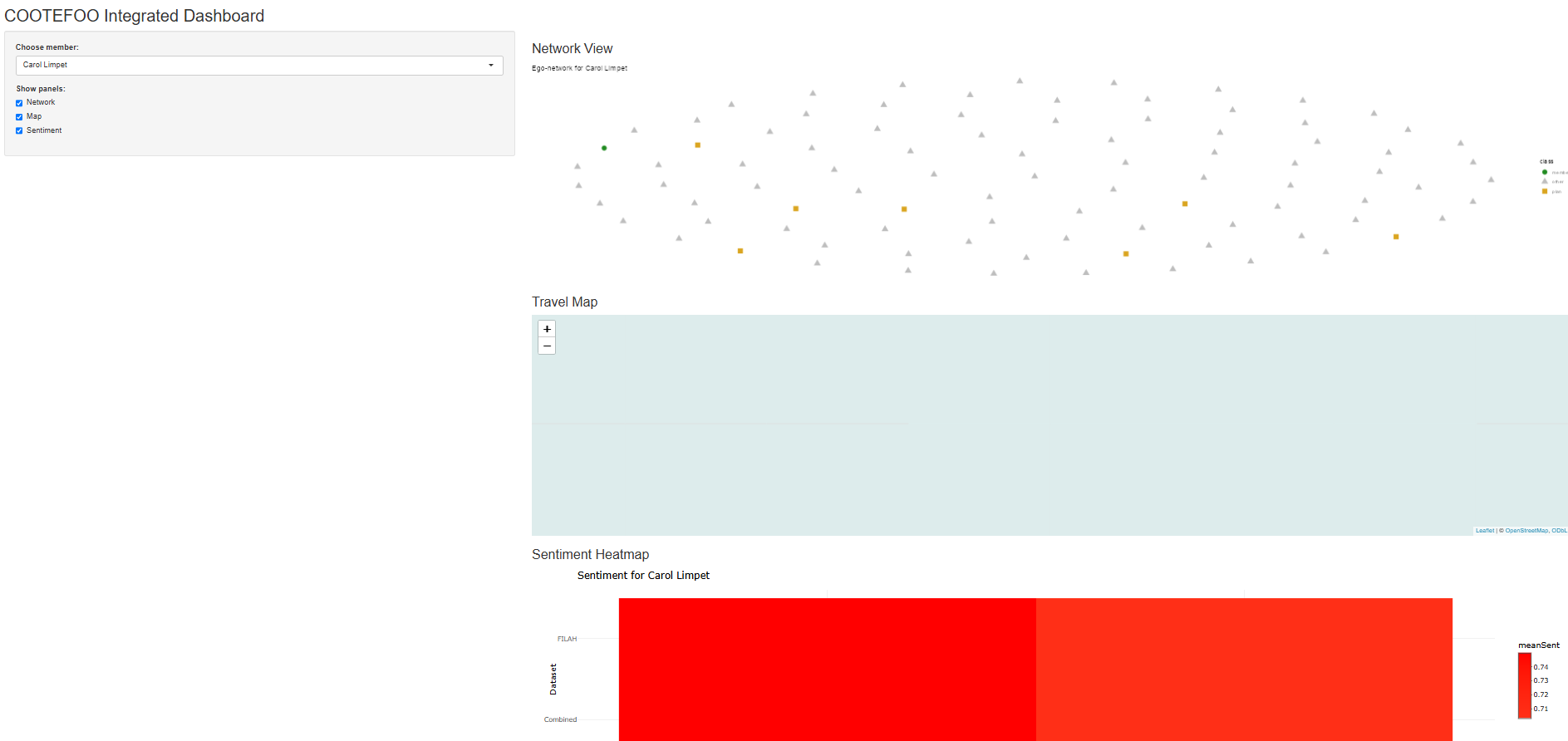
### Insights and Links to Final Stage:

This sentiment analysis explicitly reveals how initial biases from incomplete datasets are reshaped by comprehensive evidence. Members experiencing significant sentiment shifts become key figures in the detailed personal investigations undertaken in Stage 4.

## Stage 4: Integrated Dashboard - Deep Dive into Individual Members and Evidence Gaps

Finally, Moray needs precise, personalised analysis, clearly seeing how specific members’ reputations are influenced by missing or incomplete evidence. The integrated dashboard consolidates insights from previous stages, explicitly highlighting how datasets from TROUT and FILAH distorted perceptions.

### Visualisation:



An integrated, interactive dashboard explicitly combines network engagement (Stage 1), travel locations (Stage 2), and sentiment (Stage 3) into a cohesive investigative tool.

Moray selects any COOTEFOO member (defaulting explicitly to one TROUT-accused member, e.g., Seal) to synchronously update visualisations:

Ego-centric network graph explicitly visualises relationships and missing evidence.

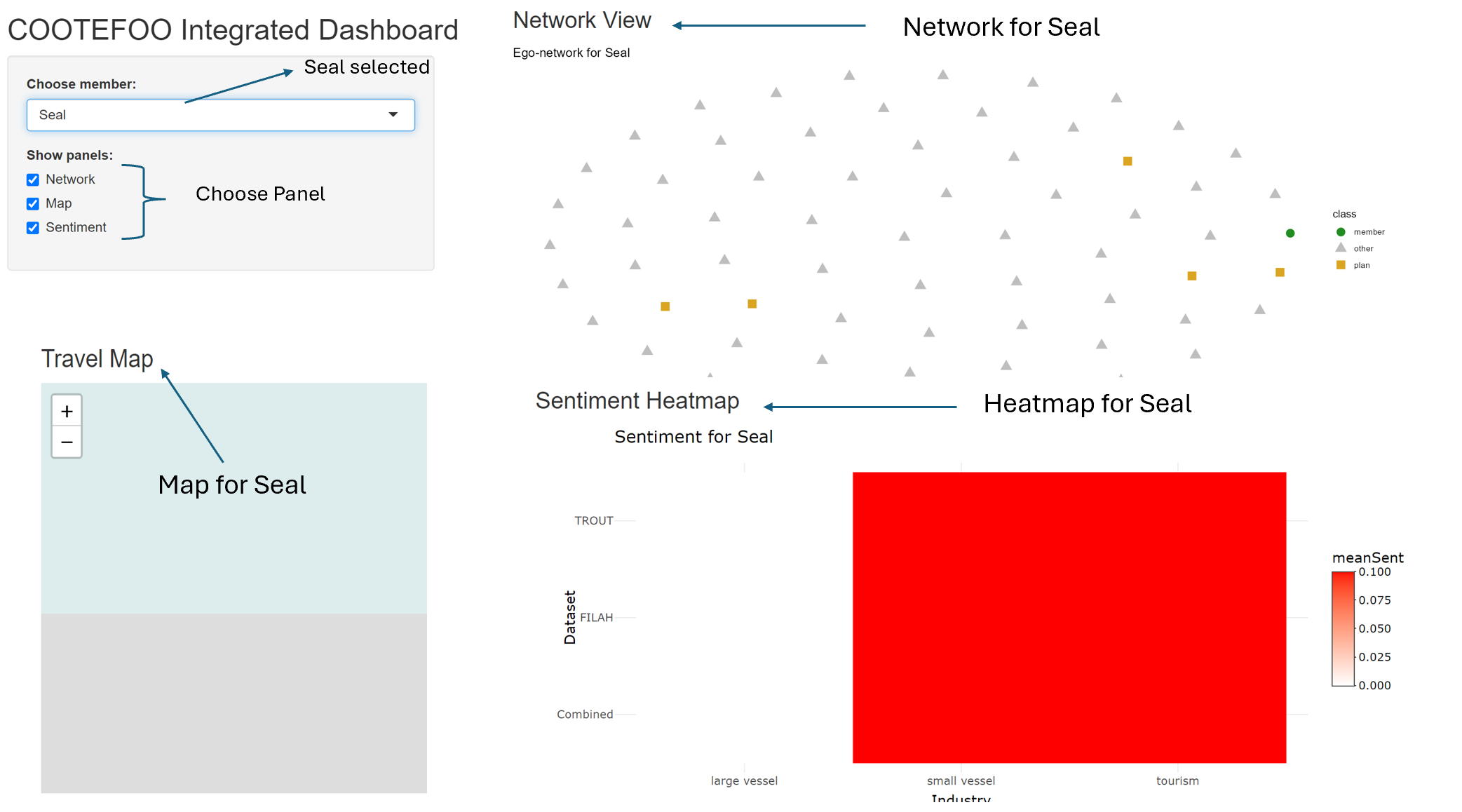
Interactive map explicitly shows detailed travel history, highlighting trips missing from individual advocacy datasets.

Sentiment heat-bar explicitly illustrates sentiment shifts between datasets, clearly quantifying bias distortions.

Detailed evidence tables explicitly list exact records (plans, discussions, trips) omitted from TROUT/FILAH, clearly revealing critical missing evidence that significantly impacts judgements.

### Illustrative Example (Seal, accused by TROUT):

Moray explicitly observes how TROUT’s incomplete evidence falsely portrays Seal as neutral. The combined data clearly reveals significant additional tourism-related activities and positive sentiment, explicitly undermining TROUT’s accusation and demonstrating precisely how missing records distort the narrative.



## Proposed Dashboard

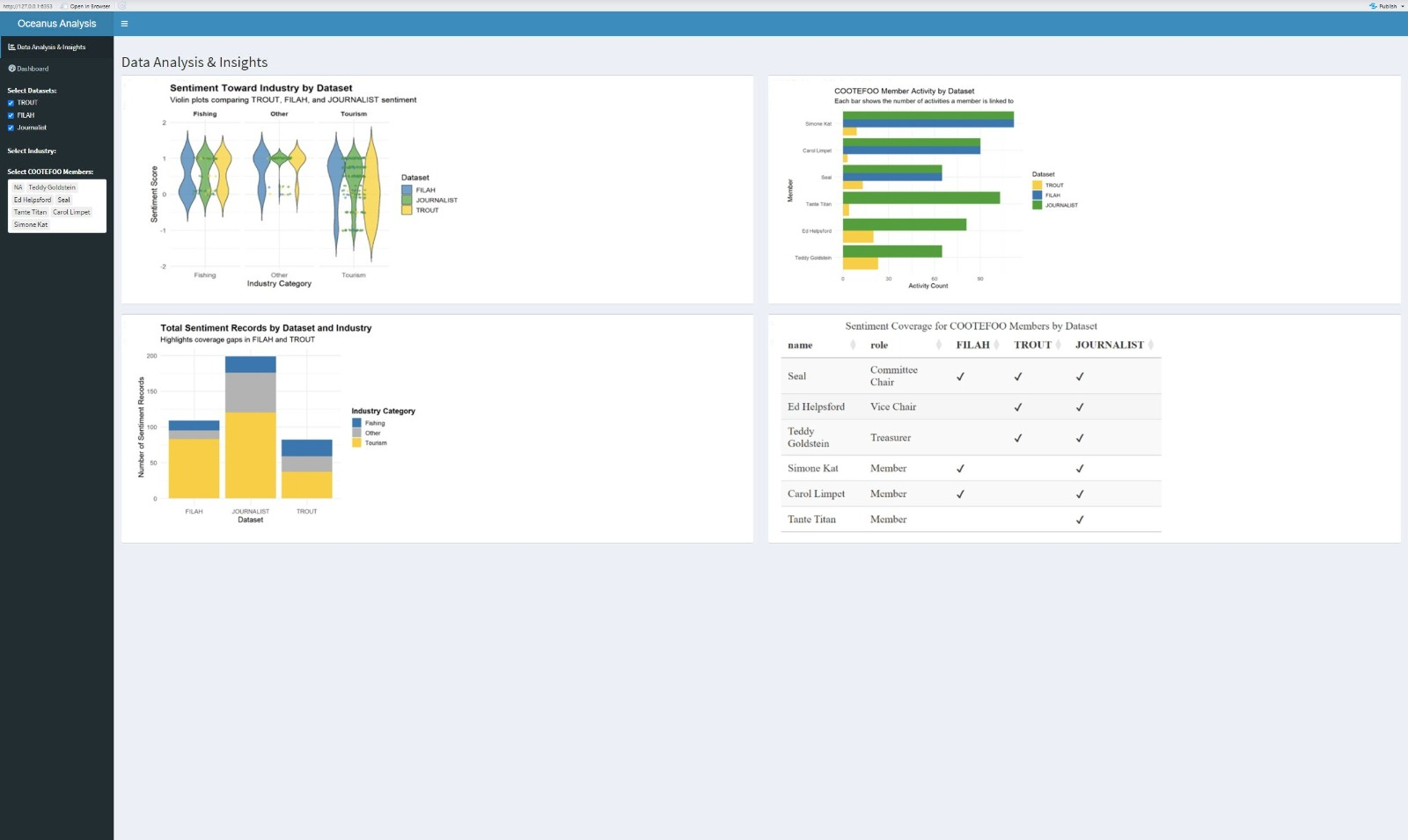
library(shiny)  
library(shinydashboard)  
library(visNetwork)  
library(leaflet)  
library(ggplot2)  
library(dplyr)  
library(tidyr)  
library(jsonlite)  
library(stringr)  
  
# Load datasets (replace with actual paths or reactiveFileReader for large files)  
datasets <- list(  
 "TROUT" = fromJSON("TROUT.json", flatten = TRUE),  
 "FILAH" = fromJSON("FILAH.json", flatten = TRUE),  
 "Journalist" = fromJSON("journalist.json", flatten = TRUE)  
)  
  
# Classify topic labels into industries  
classify\_industry\_from\_label <- function(label) {  
 if (str\_detect(label, regex("fish|dock|crane|vacuum|seafood|harbor|inspection", ignore\_case = TRUE))) {  
 return("Fishing")  
 } else if (str\_detect(label, regex("tourist|statue|park|festival|concert|market|heritage", ignore\_case = TRUE))) {  
 return("Tourism")  
 } else {  
 return("Other")  
 }  
}  
  
ui <- dashboardPage(  
 dashboardHeader(title = span("Oceanus Analysis Dashboard", style = "font-size: 20px;")),  
 dashboardSidebar(  
 sidebarMenu(  
 menuItem("Data Analysis & Insights", tabName = "insights", icon = icon("chart-bar")),  
 menuItem("Dashboard", tabName = "dashboard", icon = icon("tachometer-alt")),  
 checkboxGroupInput("selected\_datasets", "Datasets (3 data sets) - multi select:",   
 choices = names(datasets), selected = names(datasets)),  
 uiOutput("industry\_ui"),  
 uiOutput("member\_ui")  
 )  
 ),  
 dashboardBody(  
 tags$head(tags$style(HTML(".content-wrapper { overflow-x: scroll; }"))),  
 tabItems(  
 tabItem(tabName = "insights",  
 fluidPage(  
 h2("Data Analysis & Insights"),  
 fluidRow(  
 box(title = "Member Participation Count", width = 6, height = 600, plotOutput("participation\_plot", height = 550)),  
 box(title = "Distribution of Topics by Dataset", width = 6, height = 600, plotOutput("topic\_distribution\_plot", height = 550))  
 ),  
 fluidRow(  
 box(title = "Sentiment Distribution", width = 6, height = 600, plotOutput("sentiment\_distribution\_plot", height = 550)),  
 box(title = "Travel Destination Frequency", width = 6, height = 600, plotOutput("travel\_frequency\_plot", height = 550))  
 )  
 )  
 ),  
 tabItem(tabName = "dashboard",  
 fluidPage(  
 h2("Dashboard"),  
 fluidRow(  
 column(width = 6,  
 box(title = "Network Analysis of Members, Topics and Plans", width = 12, height = 600, visNetworkOutput("network\_plot", height = 550)),  
 box(title = "Geographic Analysis of Travel Patterns", width = 12, height = 600, leafletOutput("map\_plot", height = 550))  
 ),  
 column(width = 6,  
 box(title = "Sentiment Heatmap by Member and Industry", width = 12, height = 1200, plotOutput("heatmap\_plot", height = 1150))  
 )  
 )  
 )  
 )  
 )  
 )  
)  
  
server <- function(input, output, session) {  
 combined\_data <- reactive({  
 req(input$selected\_datasets)  
 bind\_rows(lapply(datasets[input$selected\_datasets], function(d) {  
 list(nodes = as\_tibble(d$nodes), edges = as\_tibble(d$edges))  
 }), .id = "source")  
 })  
   
 output$industry\_ui <- renderUI({  
 labels <- unique(unlist(lapply(datasets[input$selected\_datasets], function(d) d$edges$label)))  
 classified <- sapply(labels, classify\_industry\_from\_label)  
 checkboxGroupInput("industry\_filter", "Industry - multi select:", choices = unique(classified), selected = unique(classified))  
 })  
   
 output$member\_ui <- renderUI({  
 all\_members <- unique(unlist(lapply(datasets[input$selected\_datasets], function(d) {  
 unique(d$nodes$name[d$nodes$type == "entity.person"])  
 })))  
 selectInput("member\_filter", "COOTFFOO Member - MULTI select:", choices = all\_members, selected = all\_members, multiple = TRUE)  
 })  
   
 output$network\_plot <- renderVisNetwork({  
 data <- combined\_data()  
 nodes <- bind\_rows(lapply(datasets[input$selected\_datasets], function(d) as\_tibble(d$nodes))) %>% distinct(id, .keep\_all = TRUE)  
 edges <- bind\_rows(lapply(datasets[input$selected\_datasets], function(d) as\_tibble(d$edges)))  
   
 if (!is.null(input$industry\_filter)) {  
 edges$industry\_class <- sapply(edges$label, classify\_industry\_from\_label)  
 edges <- edges %>% filter(industry\_class %in% input$industry\_filter | is.na(industry\_class))  
 }  
   
 if (!is.null(input$member\_filter)) {  
 nodes <- nodes %>% filter(!(type == "entity.person") | name %in% input$member\_filter)  
 }  
   
 visNetwork(nodes, edges) %>% visOptions(highlightNearest = TRUE)  
 })  
   
 output$map\_plot <- renderLeaflet({  
 data <- combined\_data()  
 travel\_places <- data$nodes %>% filter(type == "place")  
 leaflet(travel\_places) %>%  
 addTiles() %>%  
 addCircleMarkers(~lon, ~lat, label = ~label, radius = 5, color = "blue", fillOpacity = 0.7)  
 })  
   
 output$heatmap\_plot <- renderPlot({  
 data <- bind\_rows(lapply(datasets[input$selected\_datasets], function(d) as\_tibble(d$edges)))  
 sentiment\_data <- data %>% filter(!is.na(sentiment), !is.na(label))  
 sentiment\_data$industry <- sapply(sentiment\_data$label, classify\_industry\_from\_label)  
   
 if (!is.null(input$industry\_filter)) {  
 sentiment\_data <- sentiment\_data %>% filter(industry %in% input$industry\_filter)  
 }  
   
 ggplot(sentiment\_data, aes(x = target, y = industry, fill = sentiment)) +  
 geom\_tile(color = "white") +  
 theme\_minimal() +  
 labs(x = "Member", y = "Industry", fill = "Sentiment") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
 })  
   
 output$participation\_plot <- renderPlot({  
 data <- bind\_rows(lapply(datasets[input$selected\_datasets], function(d) as\_tibble(d$edges)), .id = "dataset")  
 data %>% filter(!is.na(source)) %>%  
 count(dataset, source) %>%  
 ggplot(aes(x = source, y = n, fill = dataset)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Member Participation Count", x = "Member", y = "Count") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
 })  
   
 output$topic\_distribution\_plot <- renderPlot({  
 data <- bind\_rows(lapply(datasets[input$selected\_datasets], function(d) as\_tibble(d$edges)), .id = "dataset")  
 data %>% filter(type == "topic") %>%  
 mutate(industry = sapply(label, classify\_industry\_from\_label)) %>%  
 count(dataset, industry) %>%  
 ggplot(aes(x = dataset, y = n, fill = industry)) +  
 geom\_bar(stat = "identity", position = "stack") +  
 labs(title = "Topics Discussed by Industry", x = "Dataset", y = "Count")  
 })  
   
 output$sentiment\_distribution\_plot <- renderPlot({  
 data <- bind\_rows(lapply(datasets[input$selected\_datasets], function(d) as\_tibble(d$edges)), .id = "dataset")  
 data %>% filter(!is.na(sentiment)) %>%  
 ggplot(aes(x = dataset, fill = sentiment)) +  
 geom\_bar(position = "fill") +  
 labs(title = "Sentiment Distribution", y = "Proportion", x = "Dataset")  
 })  
   
 output$travel\_frequency\_plot <- renderPlot({  
 data <- bind\_rows(lapply(datasets[input$selected\_datasets], function(d) as\_tibble(d$edges)), .id = "dataset")  
 data %>% filter(type == "Travel") %>%  
 count(dataset, target) %>%  
 ggplot(aes(x = target, y = n, fill = dataset)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Travel Destination Frequency", x = "Destination", y = "Count") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
 })  
}  
  
shinyApp(ui, server)

## Dashboard: Initial Data Analysis & Insights

This dashboard provides preliminary insights into the datasets provided by FILAH, TROUT, and JOURNALIST. The user starts by selecting datasets, industries, or specific COOTEFOO members. Four visualisations summarise:

* **Member Participation Count:** Identifies active participants across datasets.
* **Distribution of Topics by Dataset:** Compares topic focus across sources
* **Sentiment Distribution:** Summarises sentiment polarity by industry.
* **Travel Destination Frequency:** Highlights common travel patterns and potential spatial biases.

These visualisations collectively offer foundational context, highlighting gaps or biases before deeper analysis.

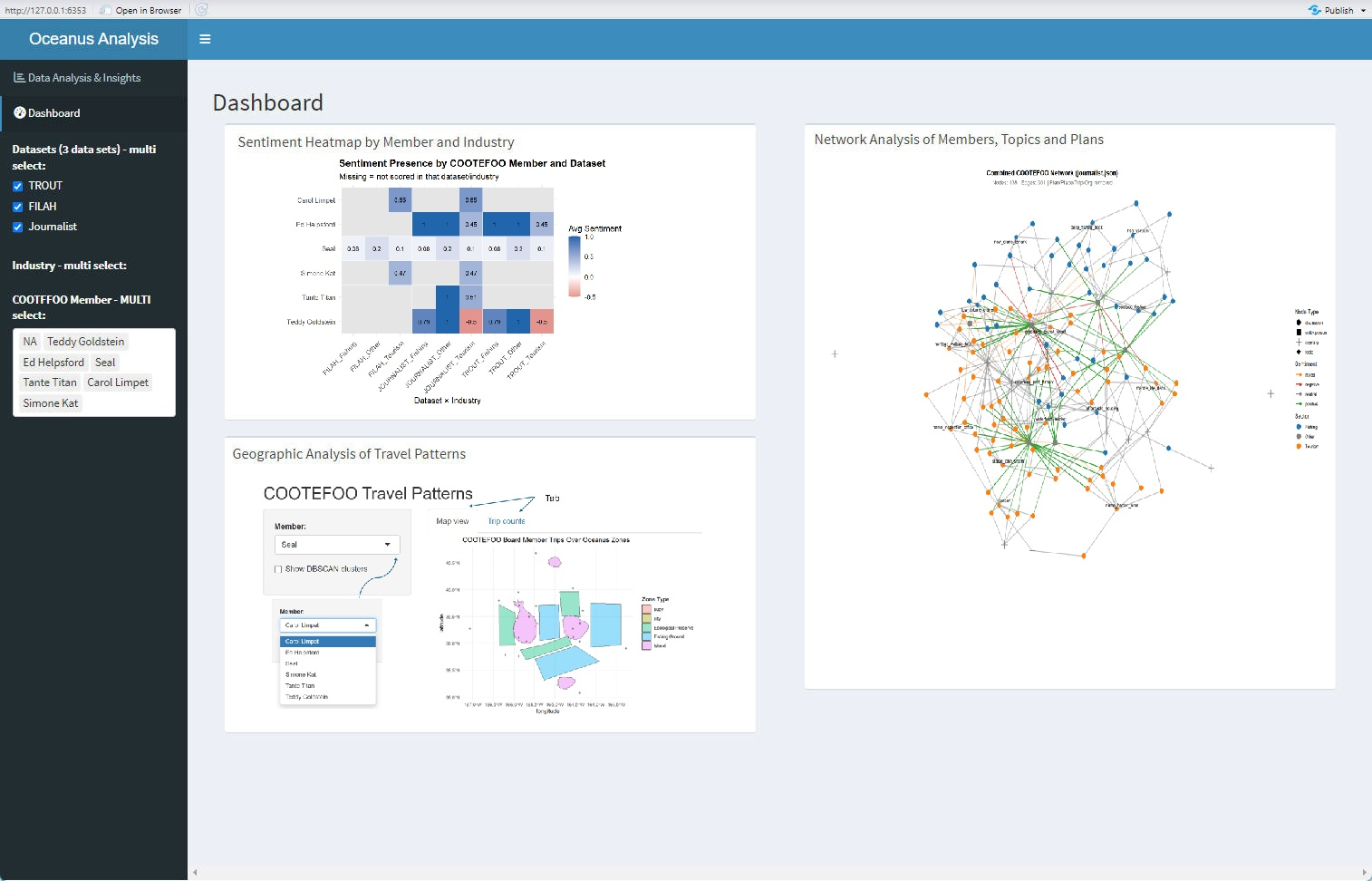


## Dashboard: Main Integrated Analysis

The main dashboard integrates sentiment, geographic, and network analysis, enabling an interactive exploration of the data:

* Sentiment Heatmap by Member and Industry: Clearly reveals biases or sentiment variations across datasets, identifying members with extreme or inconsistent stances.
* Geographic Analysis of Travel Patterns: Shows interactive geographic clustering and frequency of travel, illustrating spatial biases towards fishing or tourism.
* Network Analysis of Members, Topics, and Plans: Offers a detailed view of relationships and engagement across topics, pinpointing central figures and patterns of interactions.

The coordinated interactivity across these panels supports detailed, multifaceted exploration to robustly assess and validate claims regarding COOTEFOO’s behaviours.



# Summary

Final Integrated Summary: Original (Quarto) vs New (Shiny)

| Analytical Stage | Original Quarto Visualisation | Enhanced Shiny Interactive Visualisation |
| --- | --- | --- |
| Initial Accusations (Engagement) | Static density/bar charts | Interactive node-link network and interactive bar/dot charts |
| Spatial Bias (Travel Patterns) | Static jitter/network plots | Interactive leaflet map with clustering, filters and interactive trip counts |
| Sentiment Analysis | Single static heatmap | Interactive triple-panel heatmaps with drill-down comments and delta sentiment measures |
| Integrated Personal Analysis | Single static ego network (Seal) | Fully integrated interactive dashboard (network, map, sentiment heat-bar, detailed evidence tables), explicitly illustrating bias distortions |