DIGITAL DEMOGRAPHY: ANALYZING WEB AND SOCIAL MEDIA DATA

WEBSCRAPING - DISCUSSION

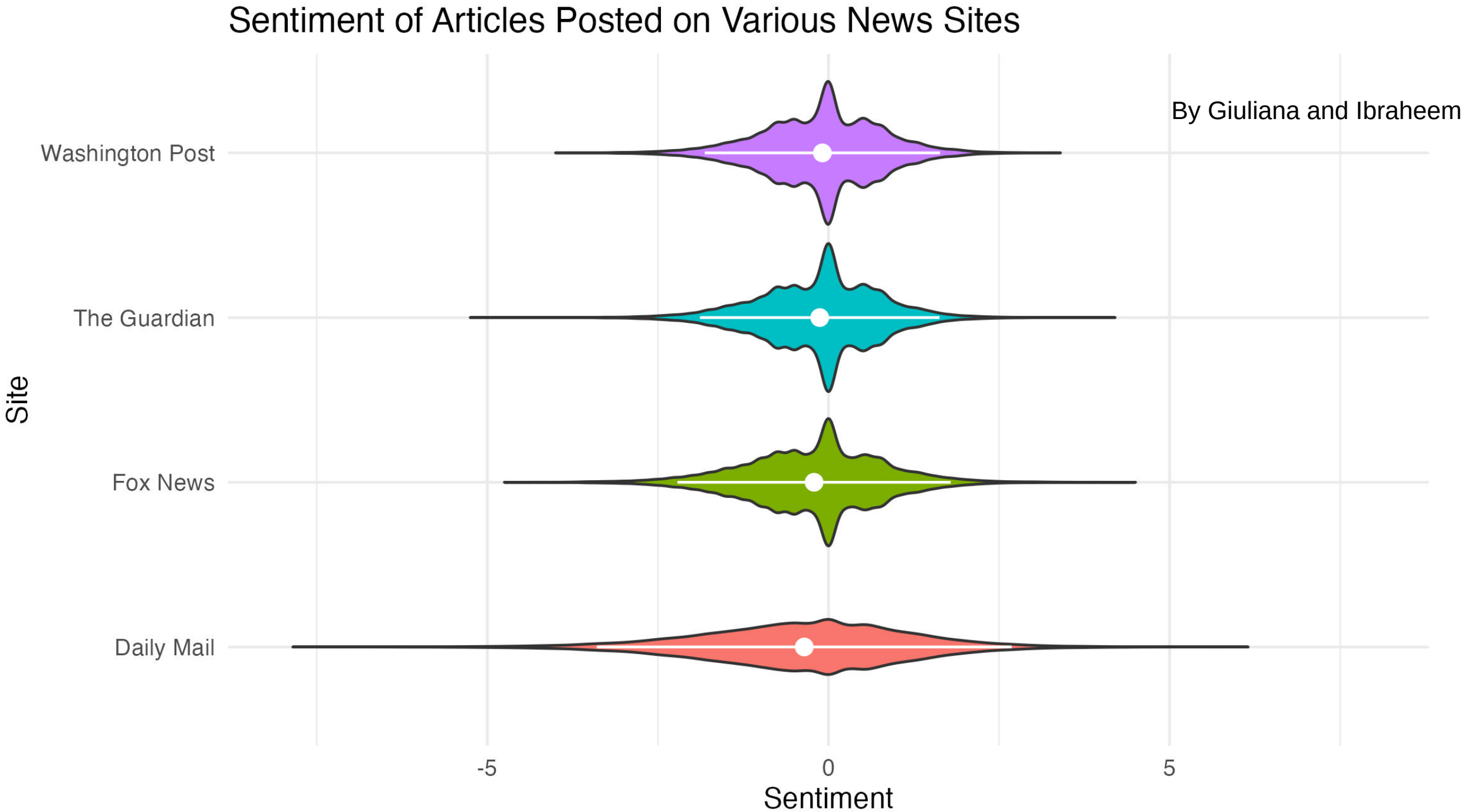
EDSD JUNE 2024

TOM THEILE

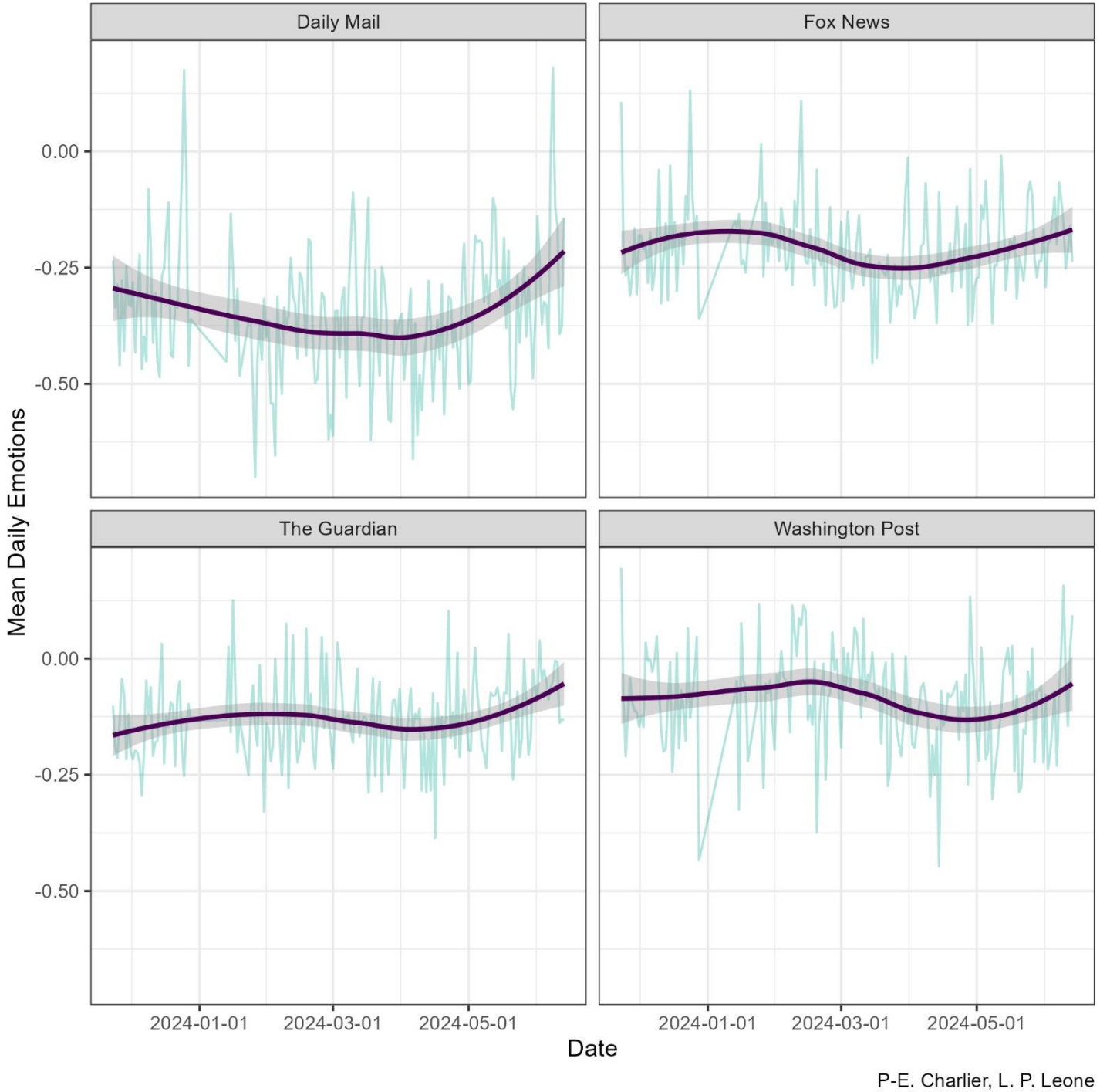
DEPARTEMENT OF DIGITAL AND COMPUTATIONAL DEMOGRAPHY

PART 2: HOW TO SCRAPE DATA FROM WEBSITES?

Your Plots!

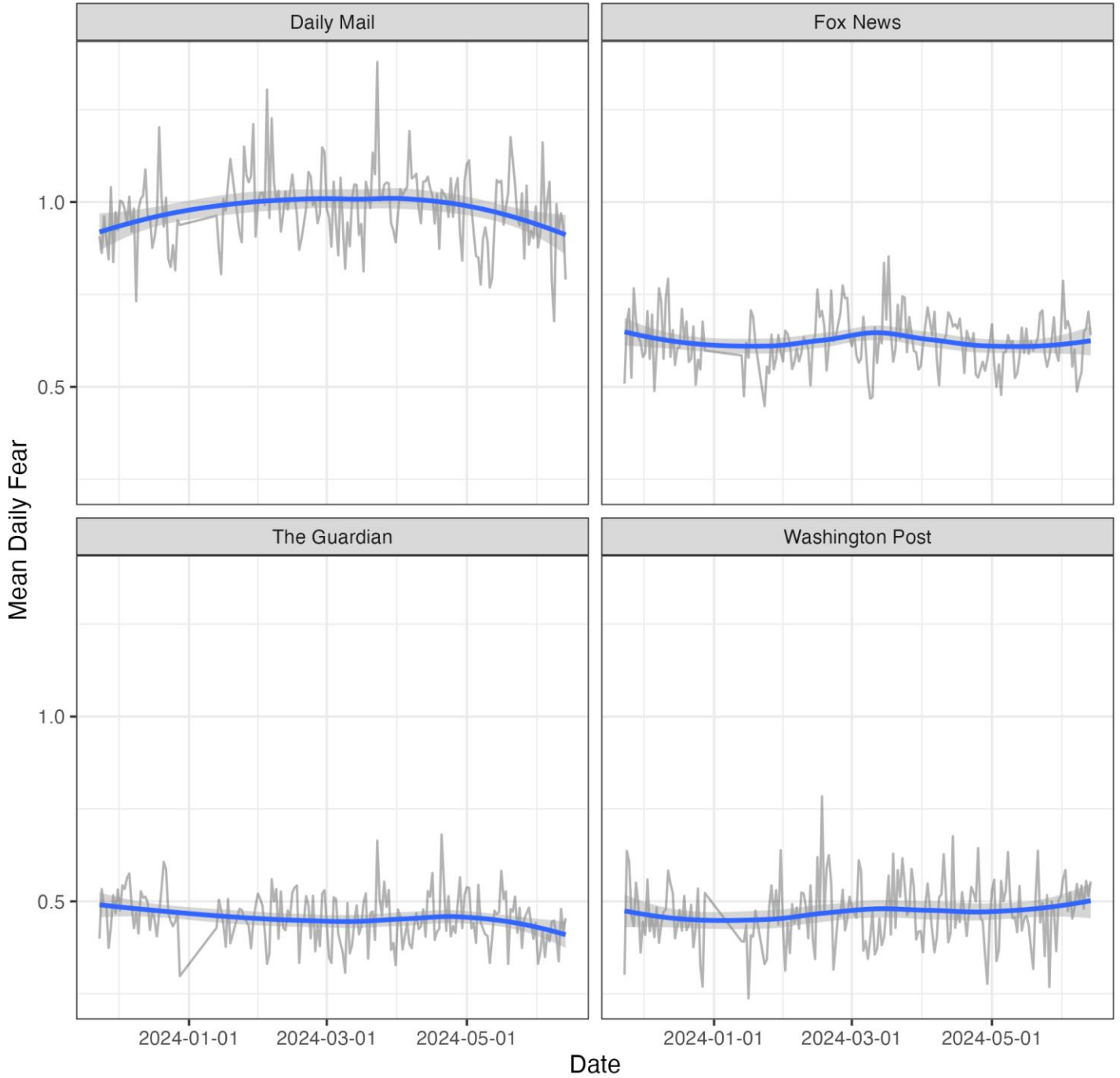

Sentiment of Articles Posted on Various News Sites
By Giuliana and Ibraheem
Washington Post
The Guardian
Fox News
Daily Mail
Site
-5
0
5
Sentiment

By Laura and Emile



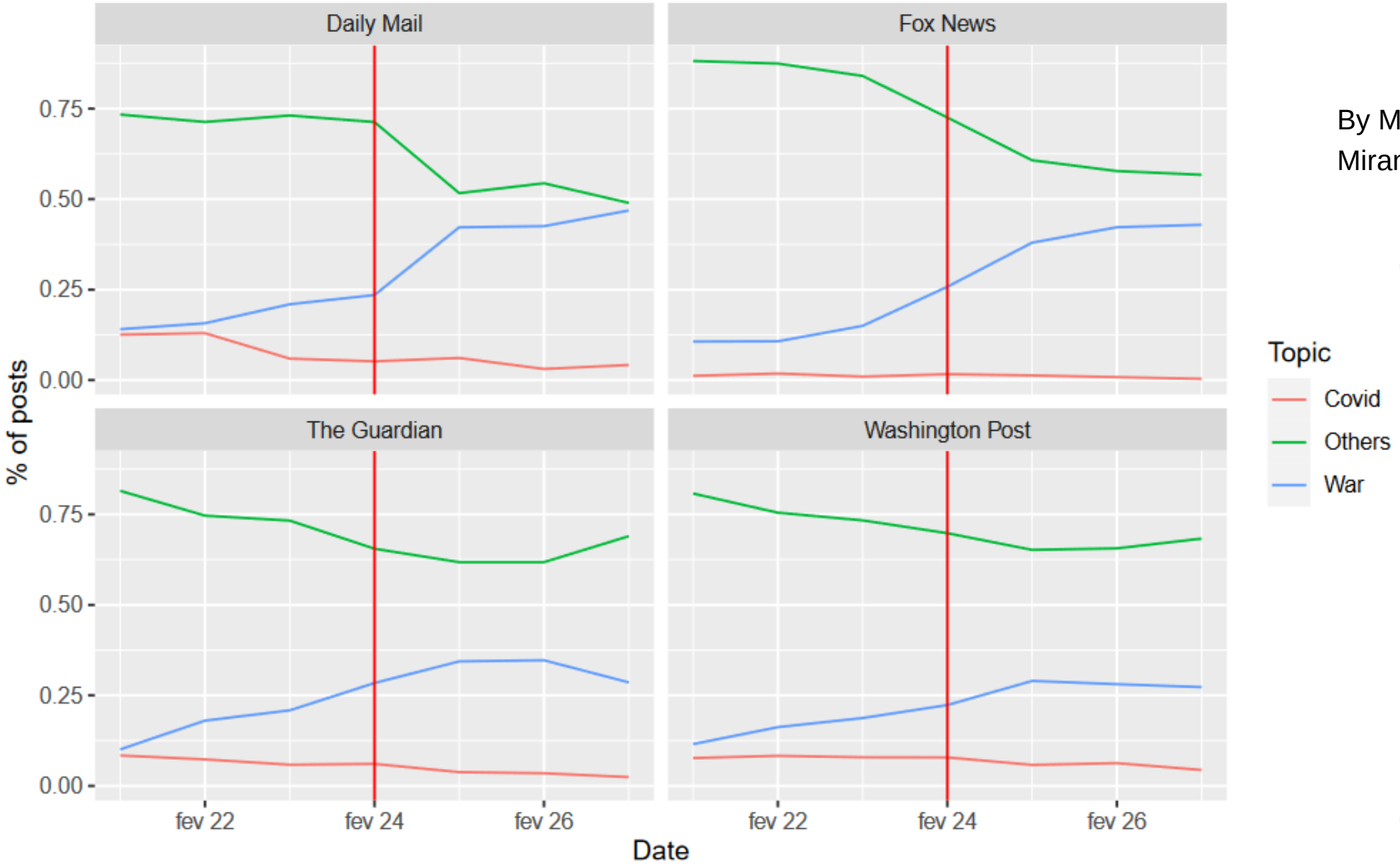
P-E. Charlier, L. P. Leone

By Laura and Emile

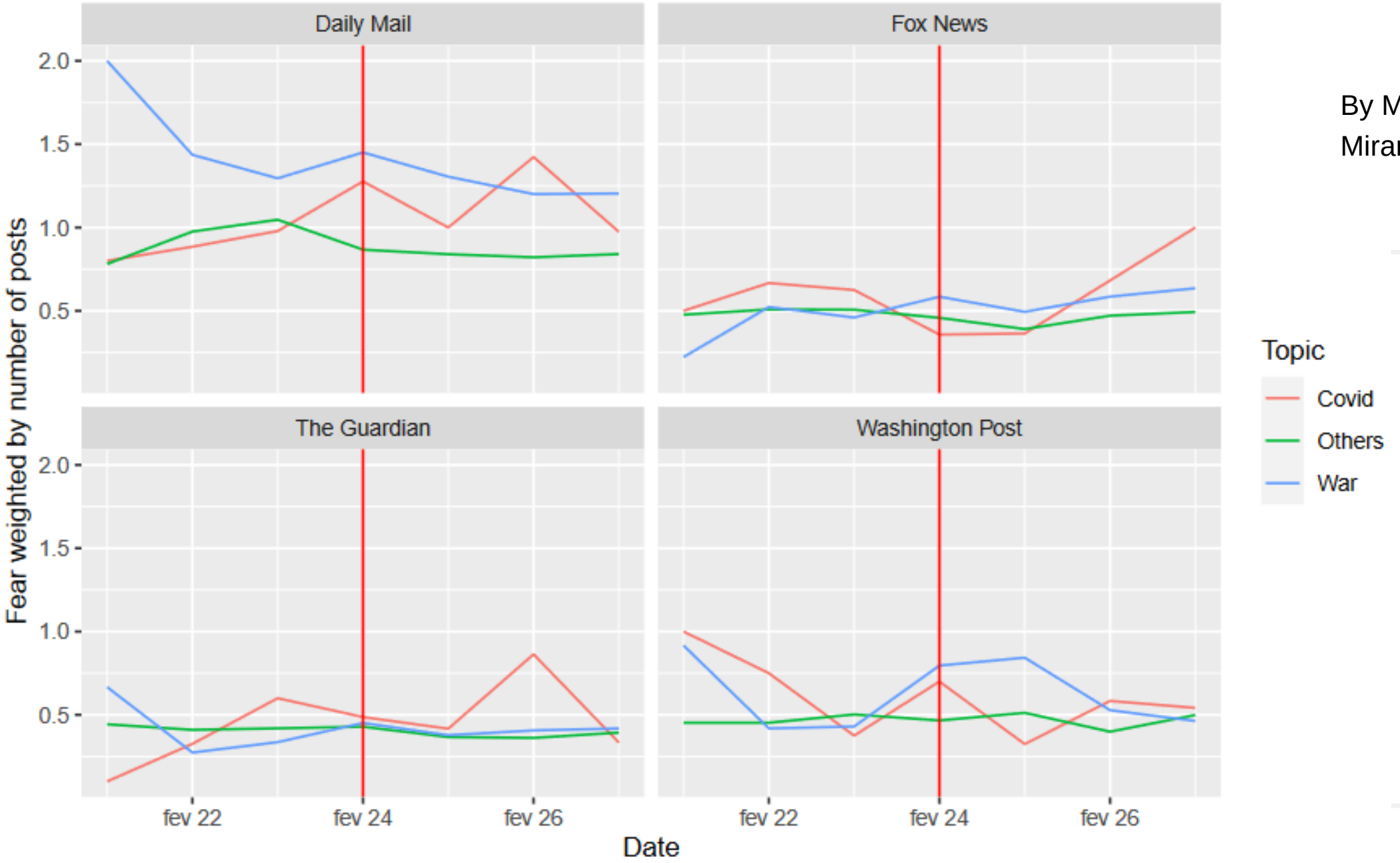


P-E. Charlier, L. P. Leone

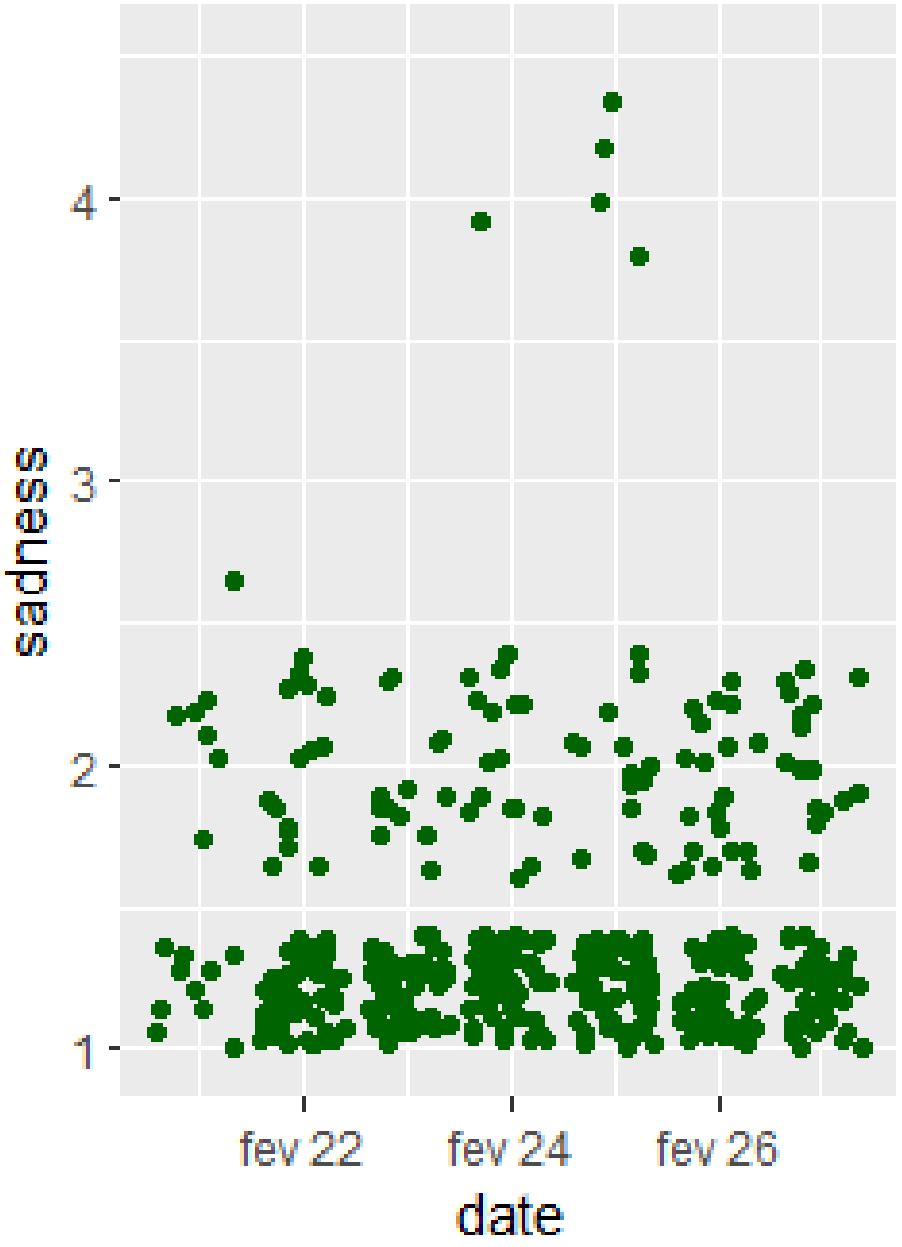
By Maria Laura Lopes Miranda

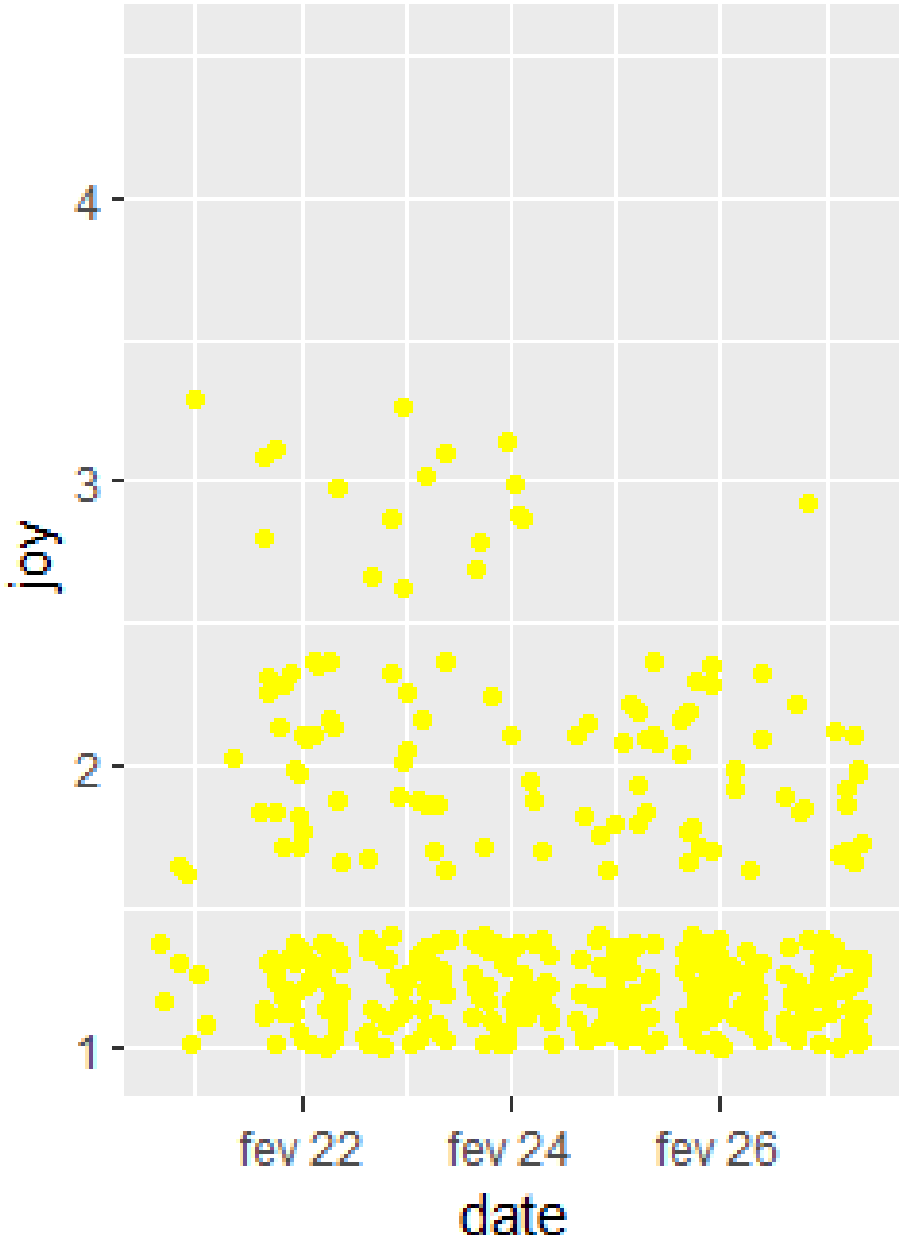


By Maria Laura Lopes Miranda



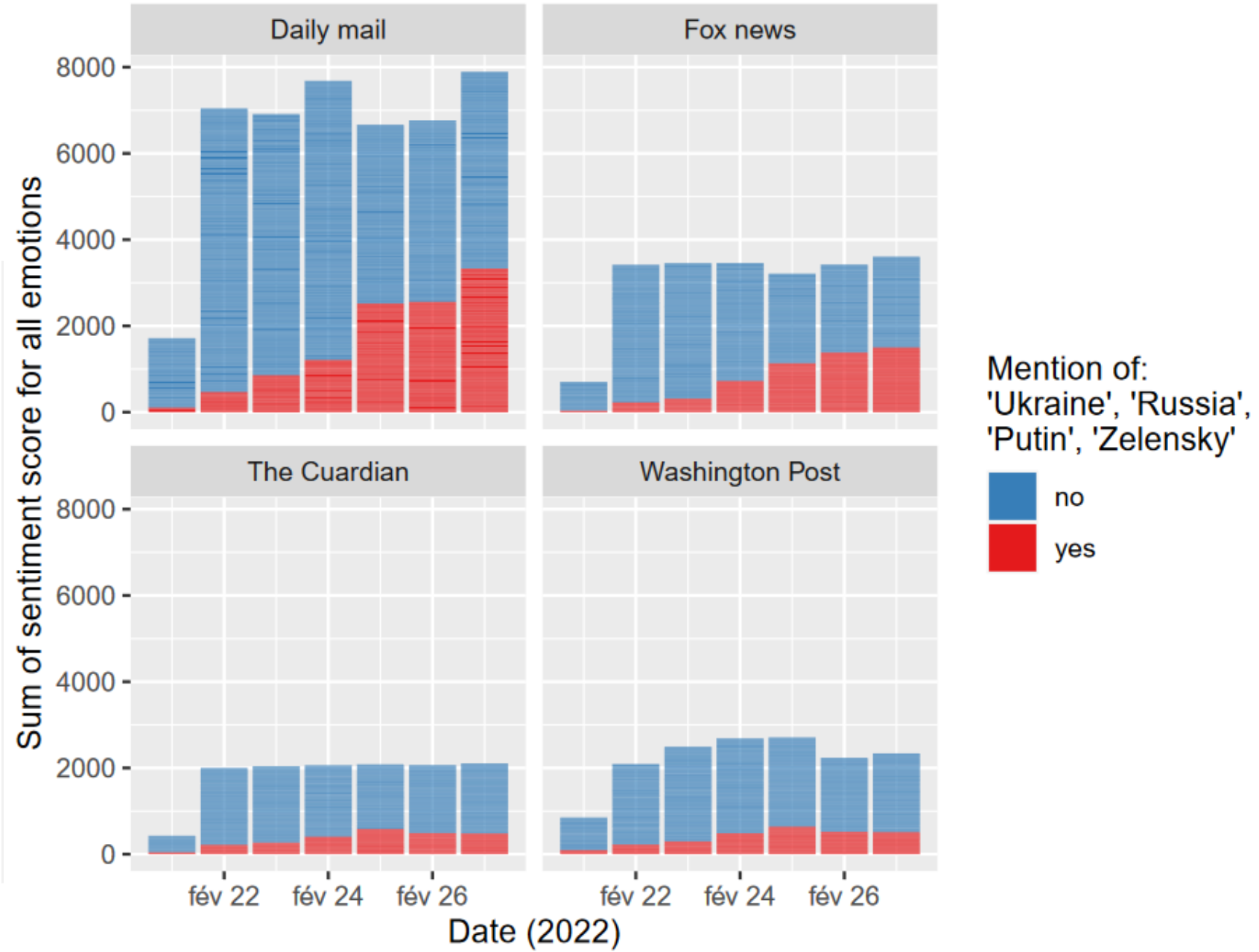
Sentimental Analysis about Brazil in The Guardian





By Amanda
Martins de Almeida

Evolution of the sentiment score during the invasion of Ukraine



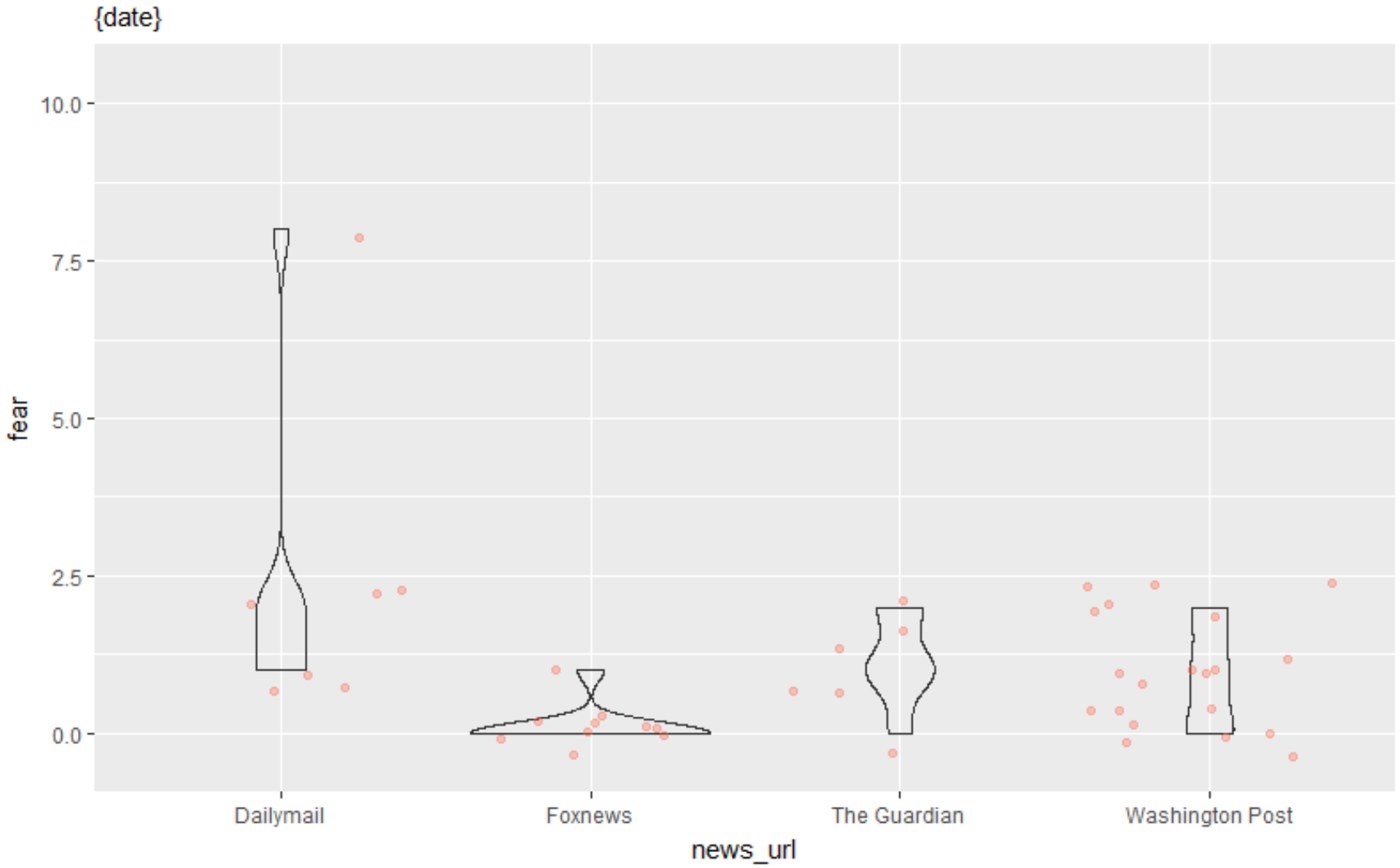
Author: Ariane Sessego, 2022 Data: webscrapping, Tim Theile, february 2022.

Evolution of the nature emotions expressed in the title

Author: Ariane Sessego, 2022. Data: webscrapping, Tim Theile, february 2022.

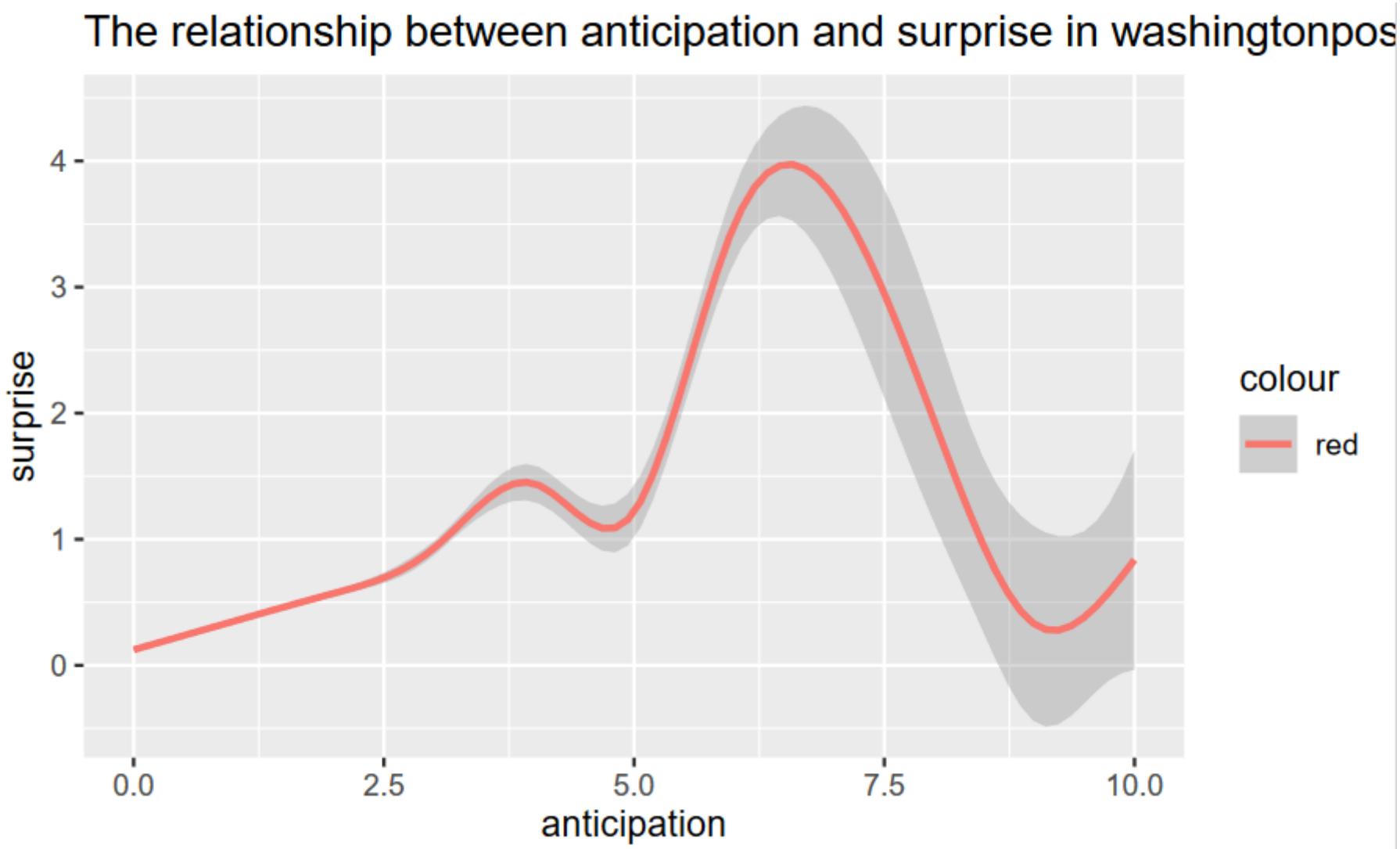
YOUR PLOTS

- Plots by Clara Girault



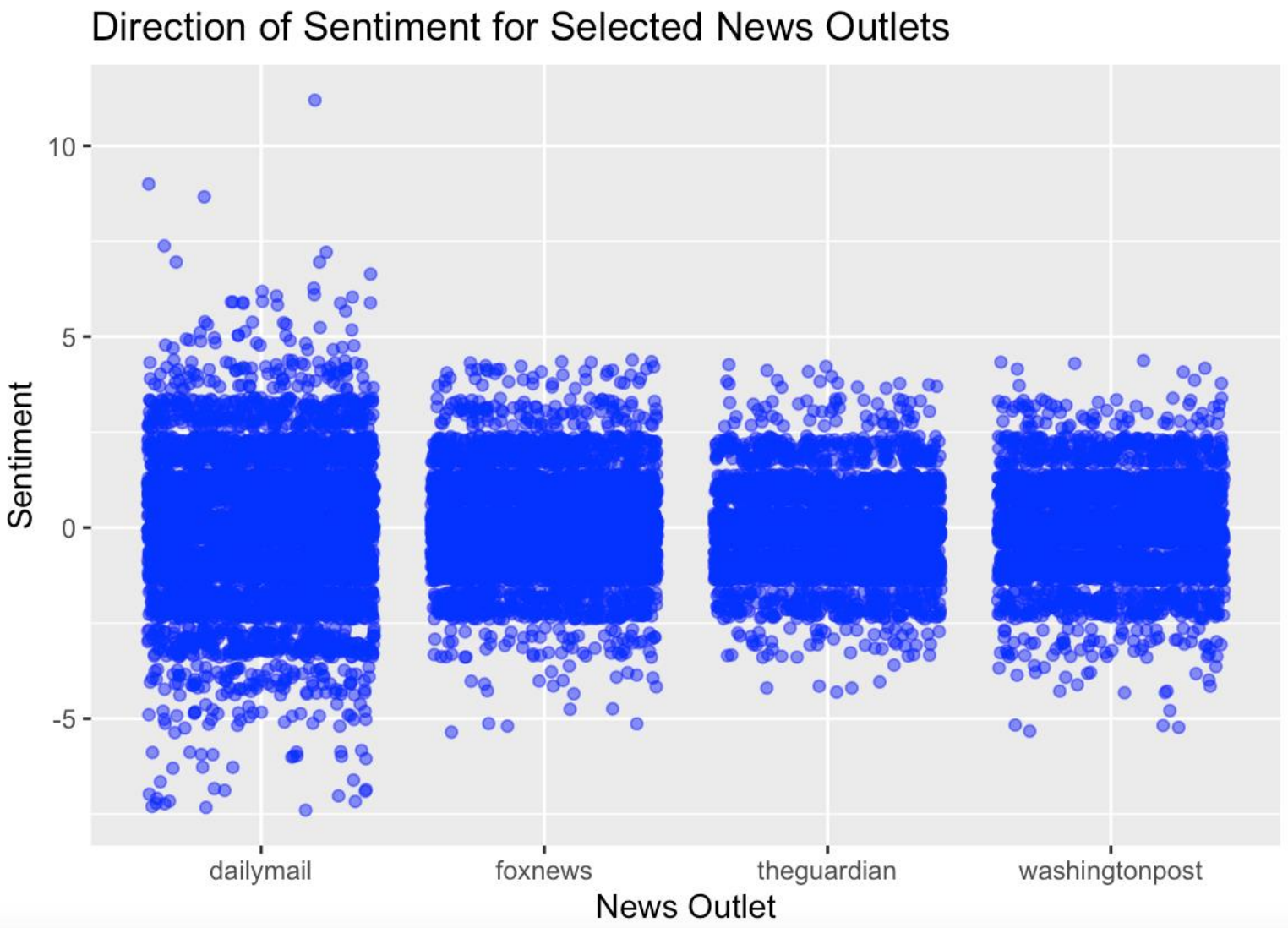
YOUR PLOTS

- Plots by Philip Orlaade



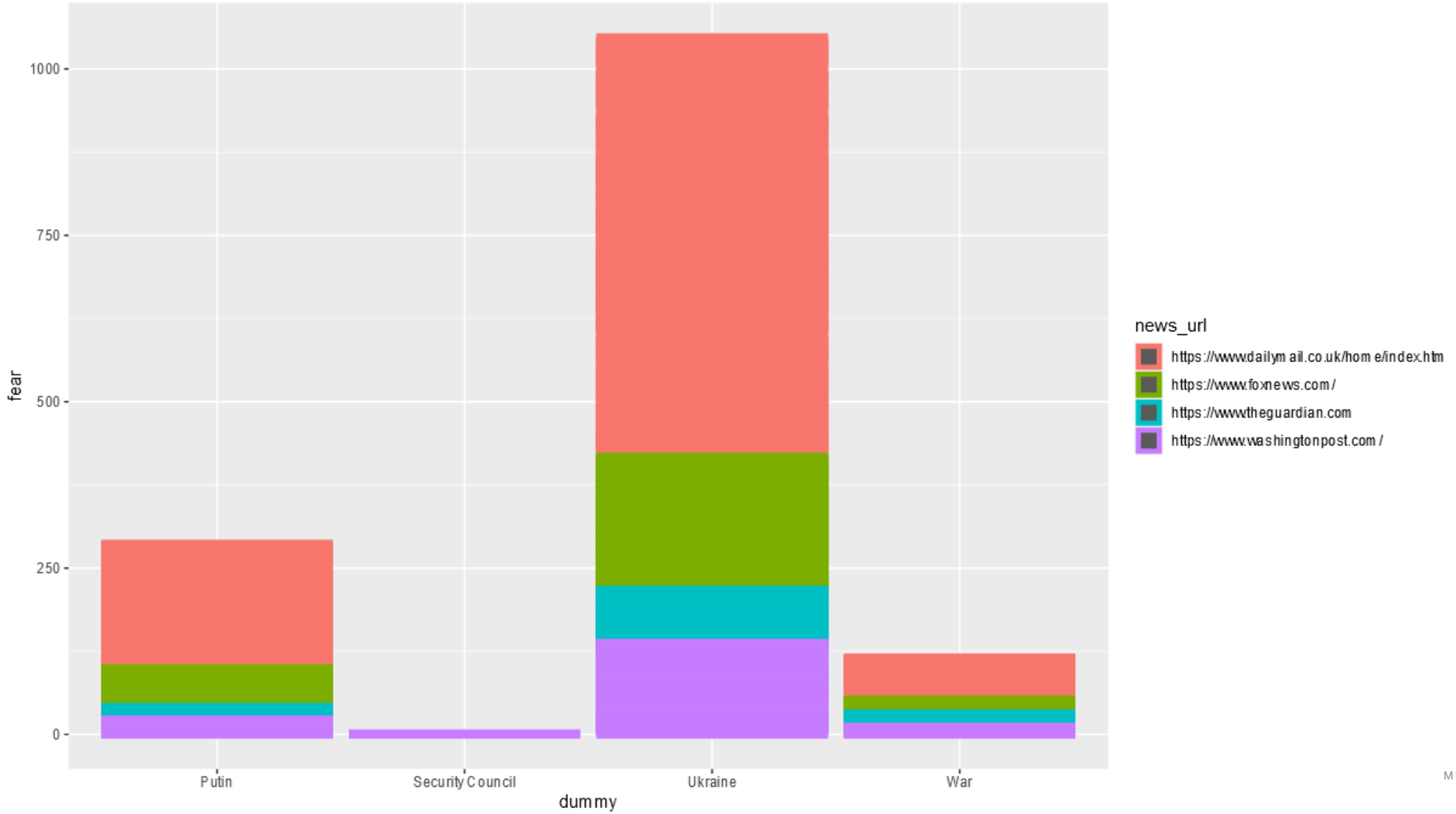
YOUR PLOTS

- Plots by Maria Louisa Christine Pohl



YOUR PLOTS

- Plots by Óskar Daði Jóhannsson

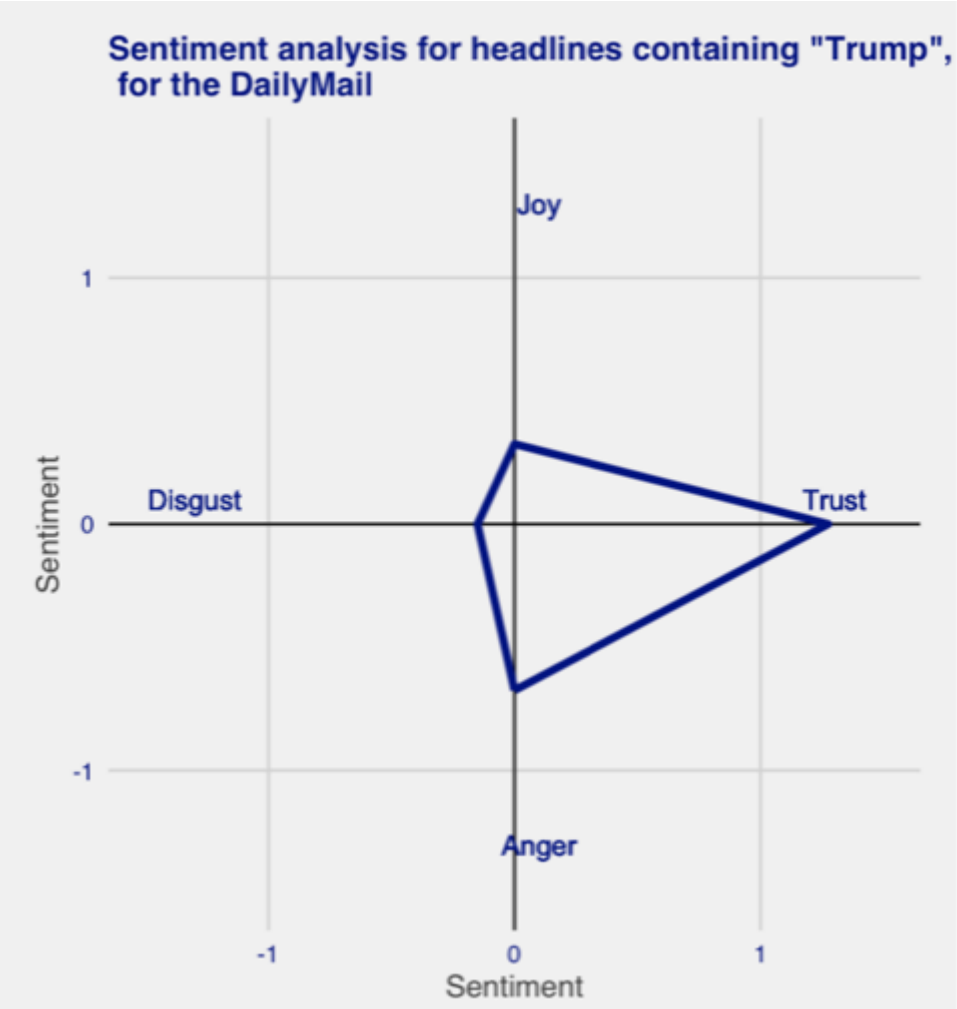


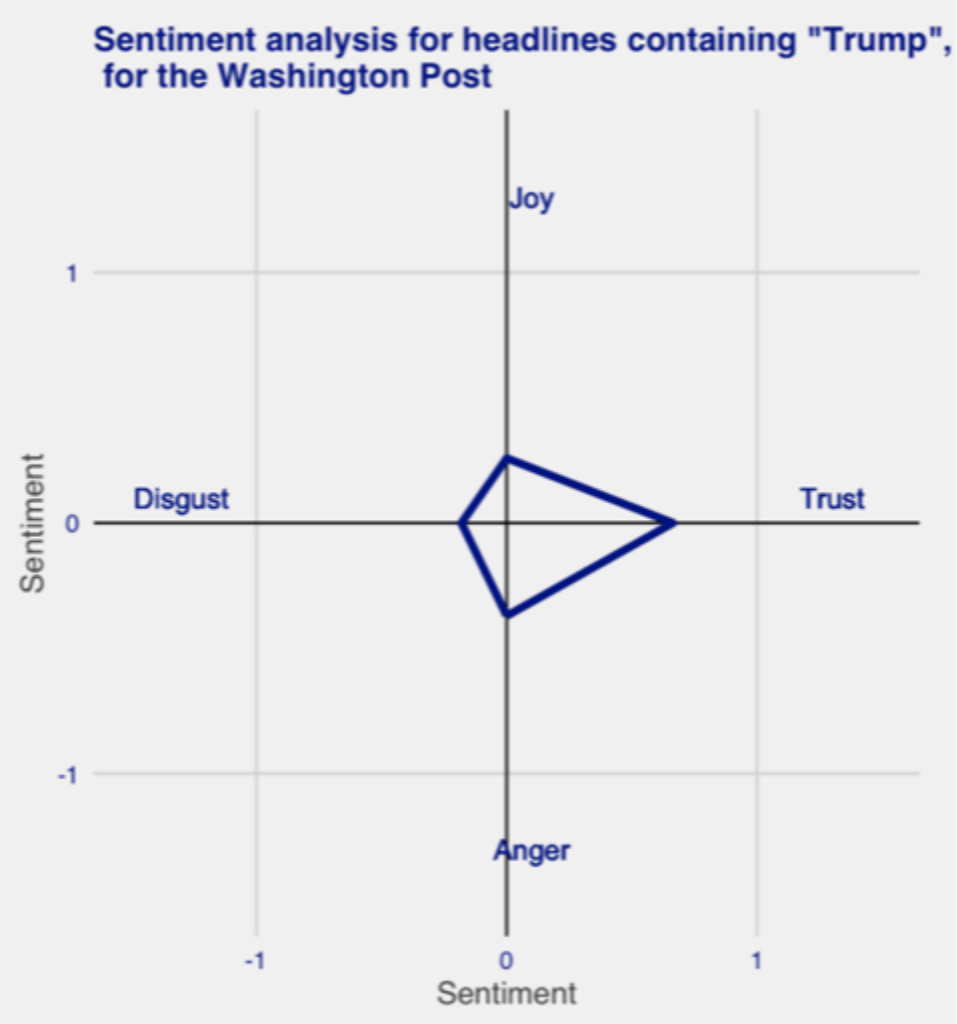
YOUR PLOTS

- Plots by

YOUR PLOTS

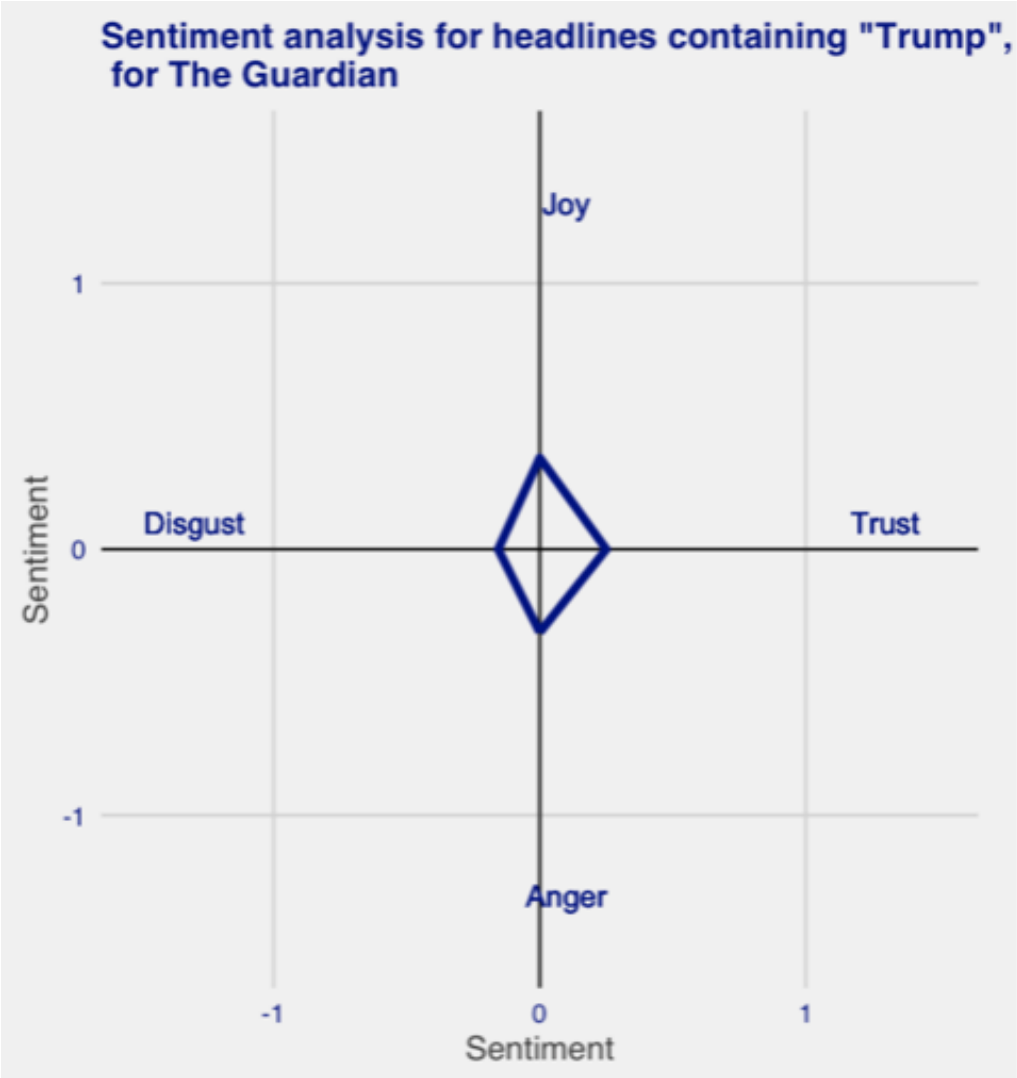
- Plots by Pietro Violo, PHDS 2022

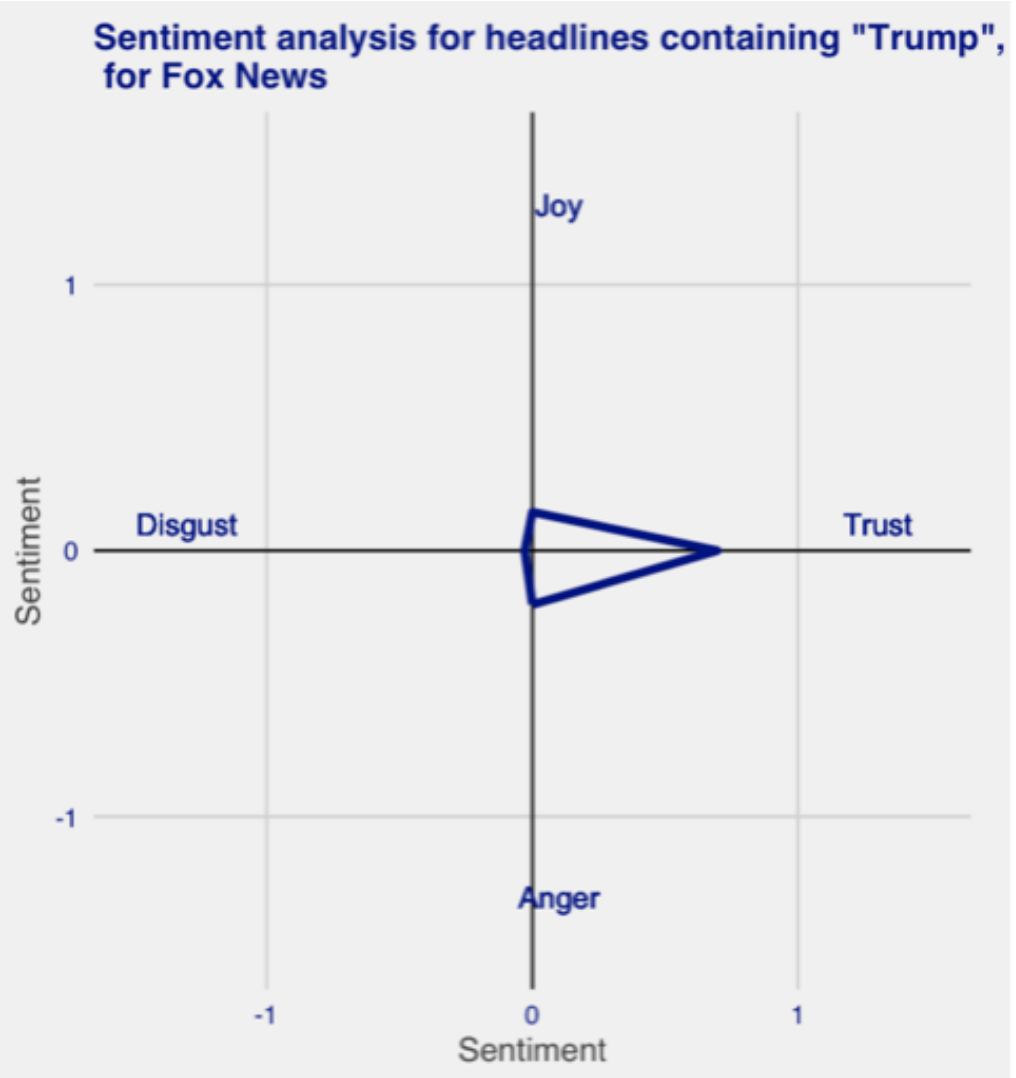




YOUR PLOTS

- Plots by Pietro Violo





PART 2: HOW TO SCRAPE DATA FROM WEBSITES?

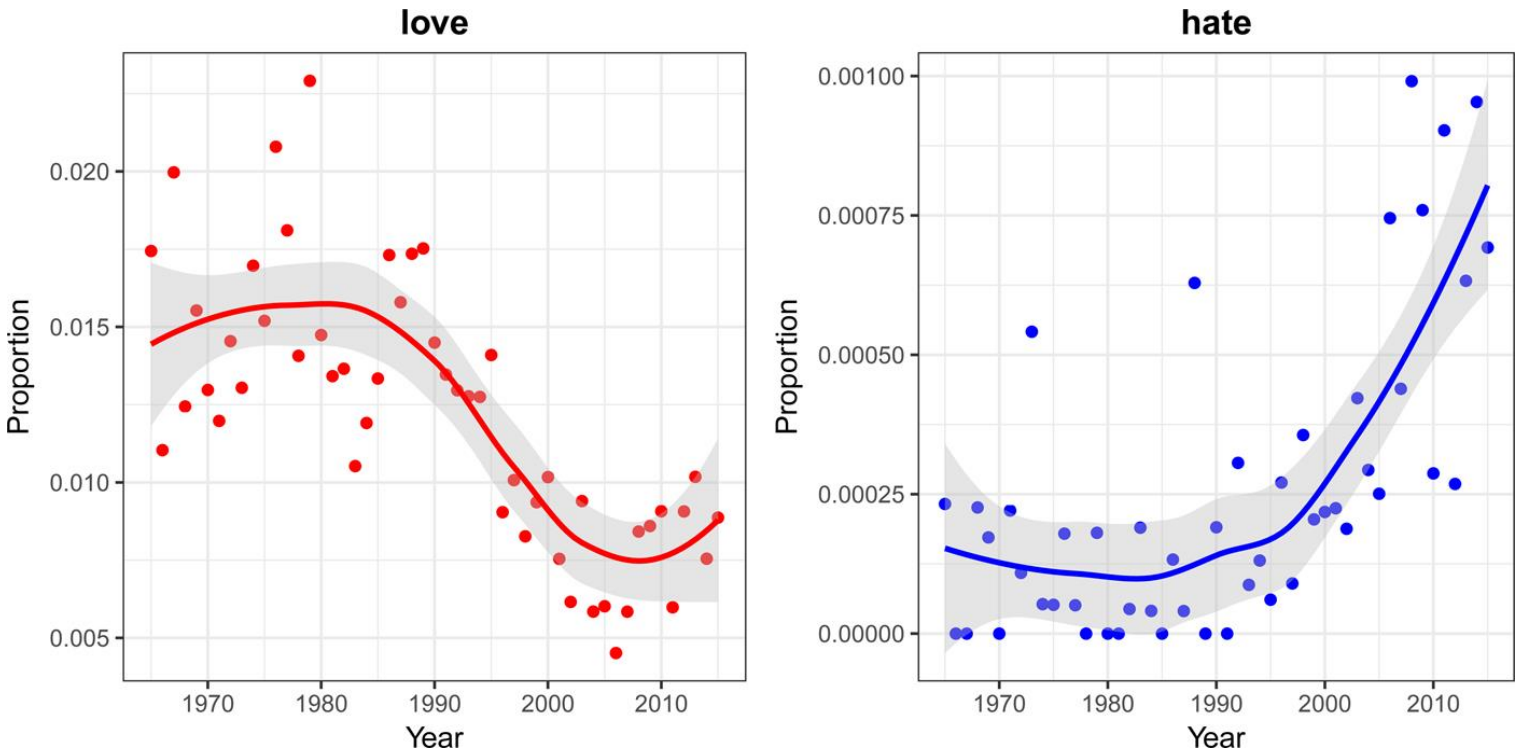
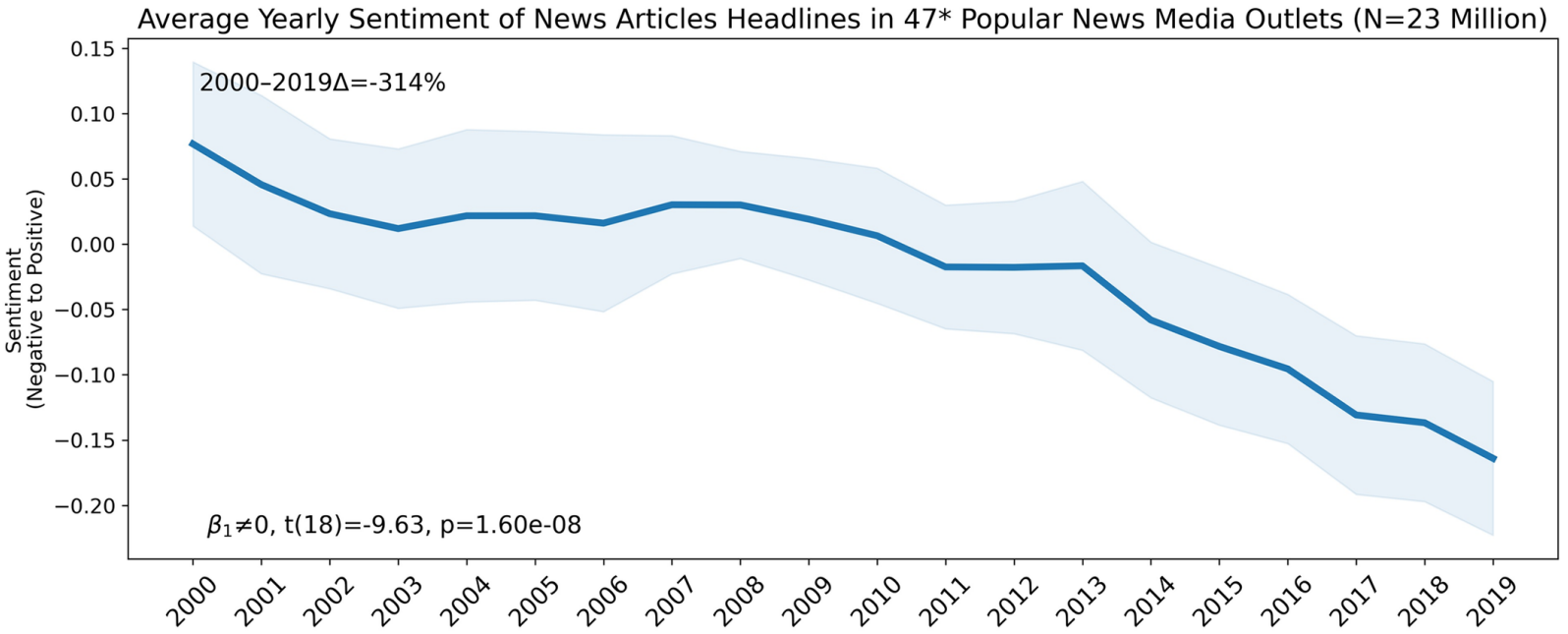


Figure 1. Proportion of the term 'love' (left panel) and 'hate' (right panel) in all song lyrics by year for the dataset billboard which contains the lyrics of the songs included in the annual US Billboard Hot 100 (n = 4913 songs). The proportions here are small as we are reporting the proportion of the word out of the total number of words in 100 songs each year (on average 30,000 words, i.e. 300 words/song) and on different scales (the frequency of positive emotion words is usually higher than the frequency of negative emotion words). To have an intuitive idea of the change, from 1965 to 1990, in the top-100 billboard songs, the word 'hate' was used each year around four or five times overall (30,000*0.00015), whereas now the average is around 24 (30,000 × 0.0008).

Cultural evolution of emotional expression in 50 years of song lyrics

Charlotte O. Brand, Alberto Acerbi and Alex Mesoudi

PART 2: HOW TO SCRAPE DATA FROM WEBSITES?

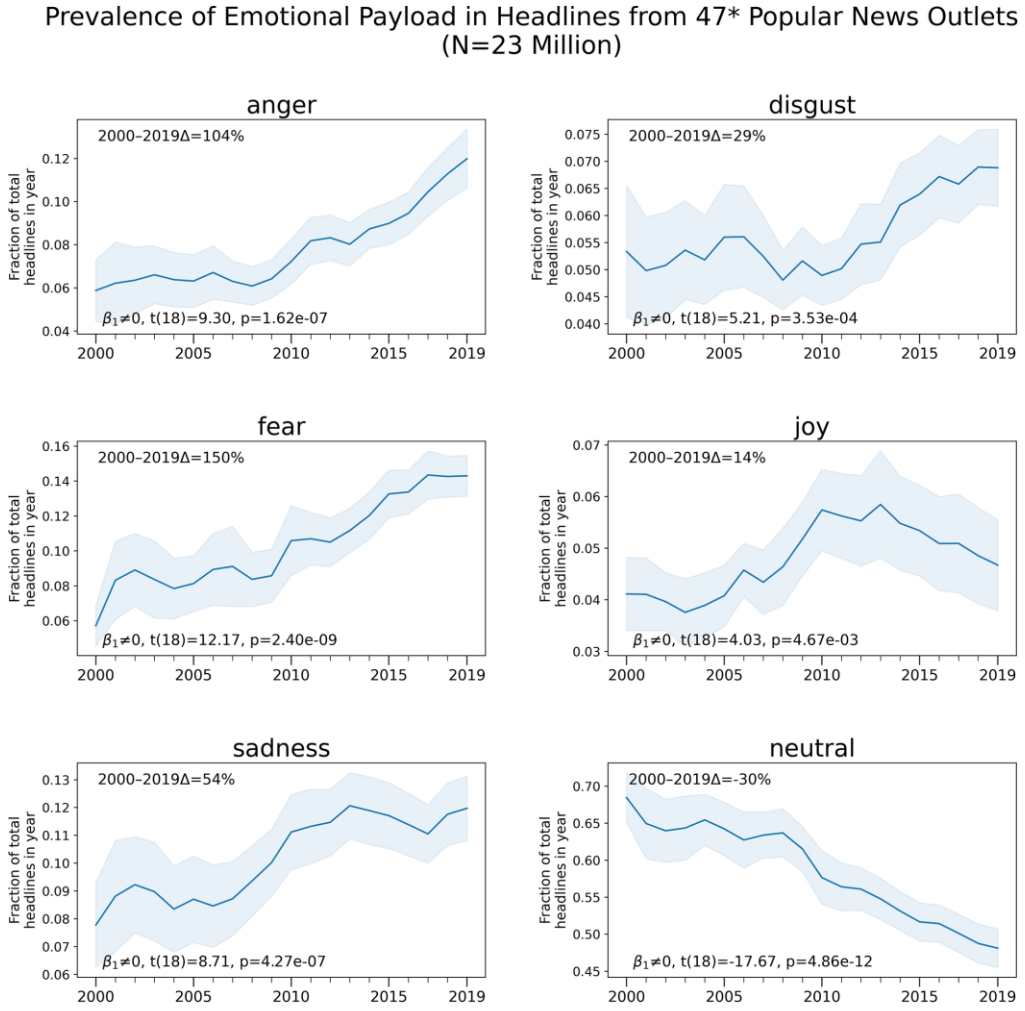


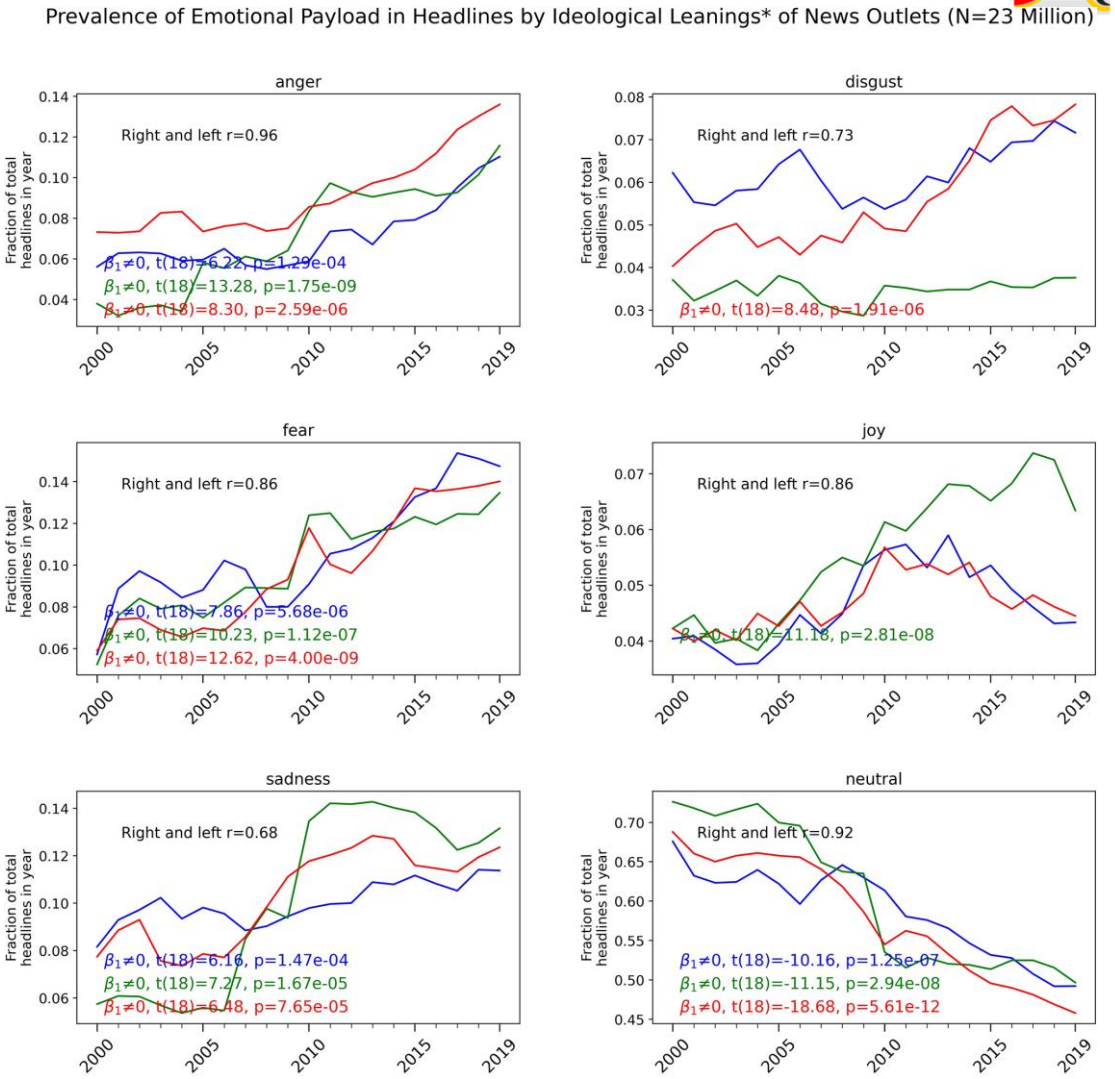
The solid blue line shows the average yearly sentiment of headlines across 47 popular news media outlets. The shaded area indicates the 95% confidence interval around the mean. A statistical test for the null hypothesis of zero slope is shown on the bottom left of the plot. The percentage change on average yearly sentiment across outlets between 2000 and 2019 is shown on the top left of the plot.

* Alternet, Democracy Now, Daily Beast, Huffington Post, The Intercept, Jacobin, Mother Jones, The New Yorker, The Nation, Slate, Vox, CNN, New York Times, ABC News, The Atlantic, Buzzfeed, CBS News, The Economist, The Guardian, NBC News, POLITICO, TIME, Washington Post, NPR, Associated Press, BBC, Bloomberg, Christian Science Monitor, REUTERS, The Hill, USA Today, Wall Street Journal, Reason, Washington Examiner, Washington Times, Fox News, American Spectator, Breitbart, The Blaze, Christian Broadcasting Network, The Daily Caller, The Daily Mail, The Daily Wire, The Federalist, National Review, New York Post, Newsmax

Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models, David Rozado , Ruth Hughes, Jamin Halberstadt, October 2022

PART 2: HOW TO SCRAPE DATA FROM WEBSITES?





* Alternet, Democracy Now, Daily Beast, Huffington Post, The Intercept, Jacobin, Mother Jones, The New Yorker, The Nation, Slate, Vox, CNN, New York Times, ABC News, The Atlantic, Buzzfeed, CBS News, The Economist, The Guardian, NBC News, POLITICO, TIME, Washington Post, NPR, Associated Press, BBC, Bloomberg, Christian Science Monitor, REUTERS, The Hill, USA Today, Wall Street Journal, Reason, Washington Examiner, Washington Times, Fox News, American Spectator, Breitbart, The Blaze, Christian Broadcasting Network, The Daily Caller, The Daily Mail, The Daily Wire, The Federalist, National Review, New York Post, Newsmax

media | **ge**

)ctober

Left leaning news outlets | Centrist news outlets | Right-leaning news outlets

Alternet, Democracy Now, Daily Beast, Huffington Post, The Intercept, Jacobin, Mother Jones, The New Yorker, The Nation, Slate, Vox, CNN, New York Times, ABC News, The Atlantic, Buzzfeed, CBS News, The Economist, The Guardian, NBC News, POLITICO, TIME, Washington Post, NPR,

Associated Press, BBC, Bloomberg, Christian Science Monitor, REUTERS, The Hill, USA Today

Wall Street Journal, Reason, Washington Examiner, Washington Times, Fox News, American Spectator, Breitbart, The Blaze, Christian Broadcasting Network, The Daily Caller, The Daily Mail,The Daily Wire, The Federalist, National Review, New York Post, Newsmax

* Political leanings labels from AllSides Media Bias Chart v1.1 (https://www.allsides.com/blog/updated-allsides-media-bias-chart-version-11)

THANK YOU FOR YOUR ATTENTION!

Tom Theile

Research Software Engineer

theile@demogr.mpg.de