Super Lexis Team

Hurdles on the road

Roeland Dubel, Dhruv Mittal, Hannah Murdock, Leevi Saari, Mariia Tepliakova, Polina Smirnova, Vera Savulescu

CCOMPIOCIATAAL
SOCIAL SCIEN CE

"Use existing databases," they said.

"It will be easy," they said.

Research topic & question

Topic: Cycling policies and infrastructure(as we are in Netherlands...)

Question: How have cycling infrastructure and policies been covered in news media outlets in terms of sentiment, topics, cities, and political orientation in a sample of European countries over time?

OUR PIPELINE (in our dreams)

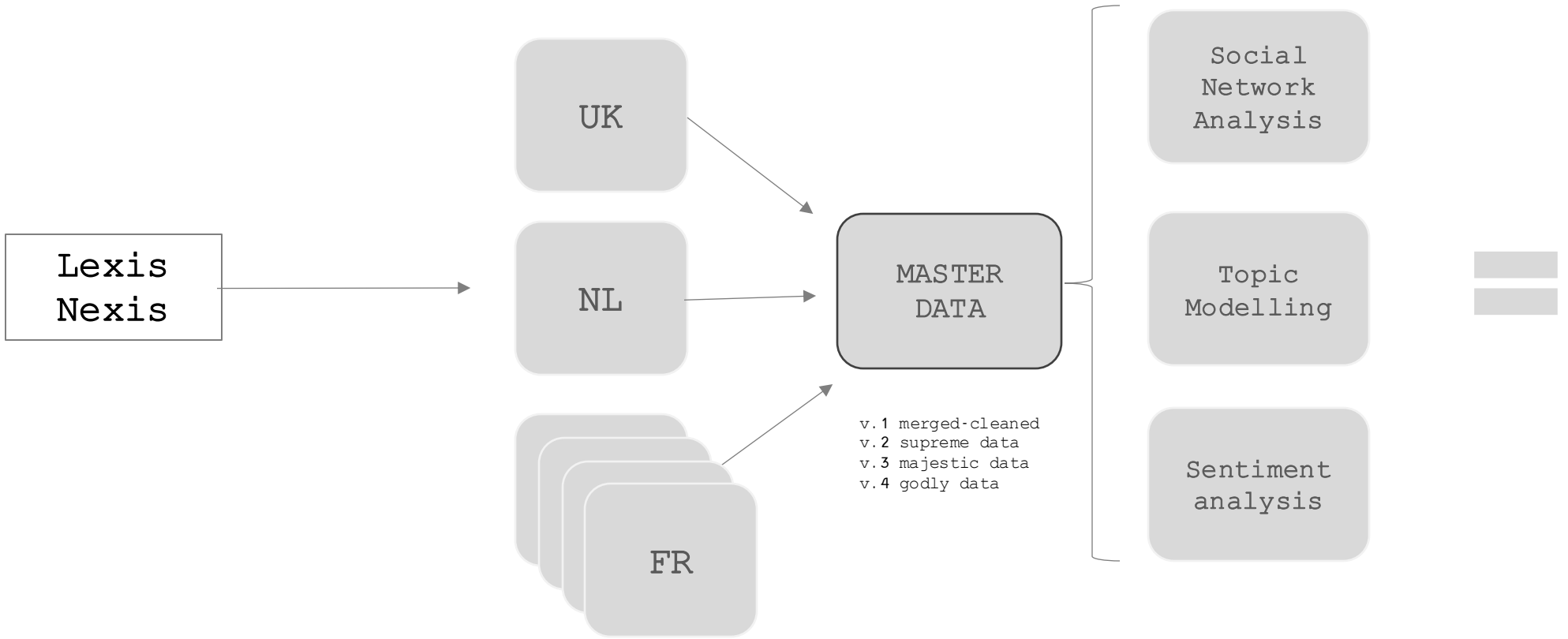
TOPIC SELECTION | IMPORT | PREPROCESS | MERGE | ANALYSIS | SCIENCE?

Parser rtf to dataframe

Translation, clean, fixing columns, post-hoc cleaning and automation

Combine to one master dataframe

Variety of methods and tools to explore our hard-won data



Many lessons through mistakes

Few interesting comparative observations

Scope

- Time frame: 2004-2024

- 2 specific news outlets in 6 European countries
  - 1 right-leaning and 1 left-leaning
  - Analogue edition over online

Country	Left-leaning	Right-leaning
France	Libération	Le Figaro
Germany	Spiegel	Frankfurter Allgemeine Zeitung (FAZ)
Spain	El País	El Mundo
Italy	La Stampa	ItaliaOggi
Netherlands	NRC Handelsblad	De Telegraaf
UK	The Guardian	The Telegraph

Data collection & cleaning

Data collection

- Search terms in 5 languages:

  Bik OR Cycl* OR Bicycle OR Velo*

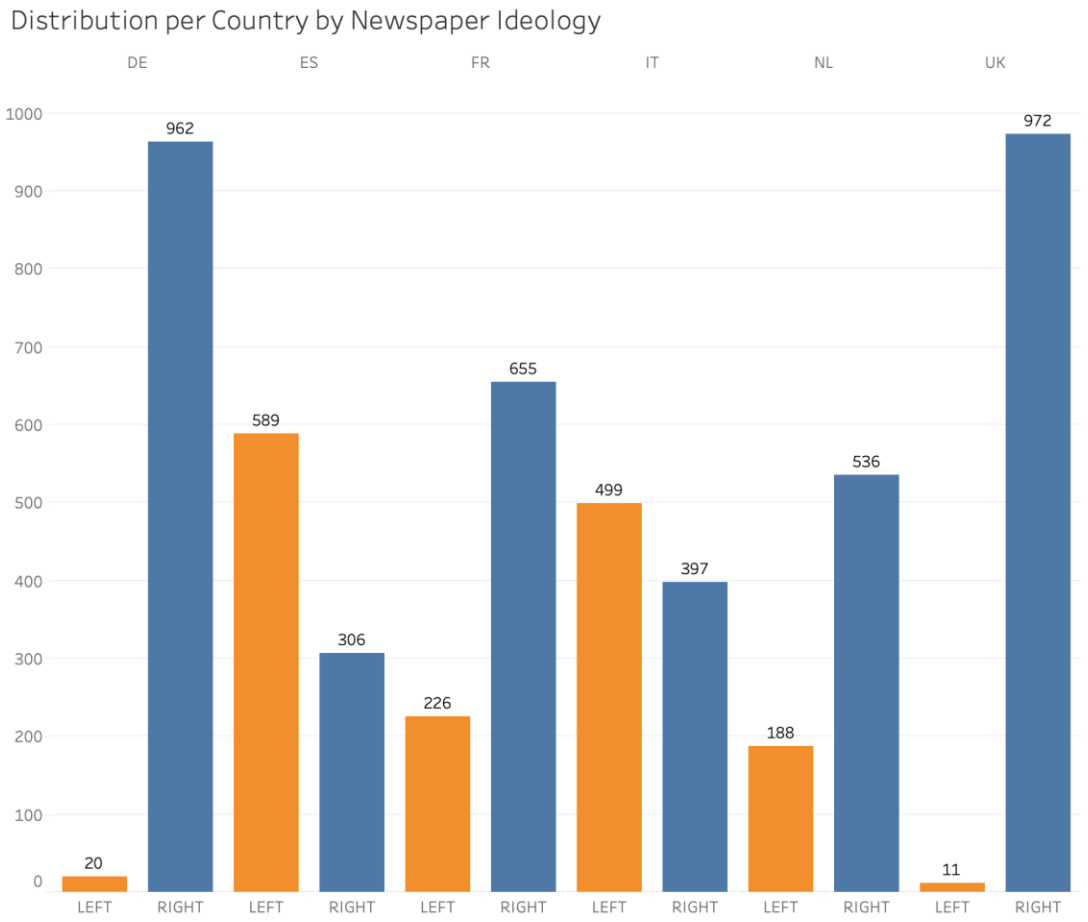
  AND
  *Infrastructure OR Routes OR Lanes OR Policies OR policy OR Initiative OR Plan OR Planning OR Strateg**
  AND
  Urban OR city OR cities

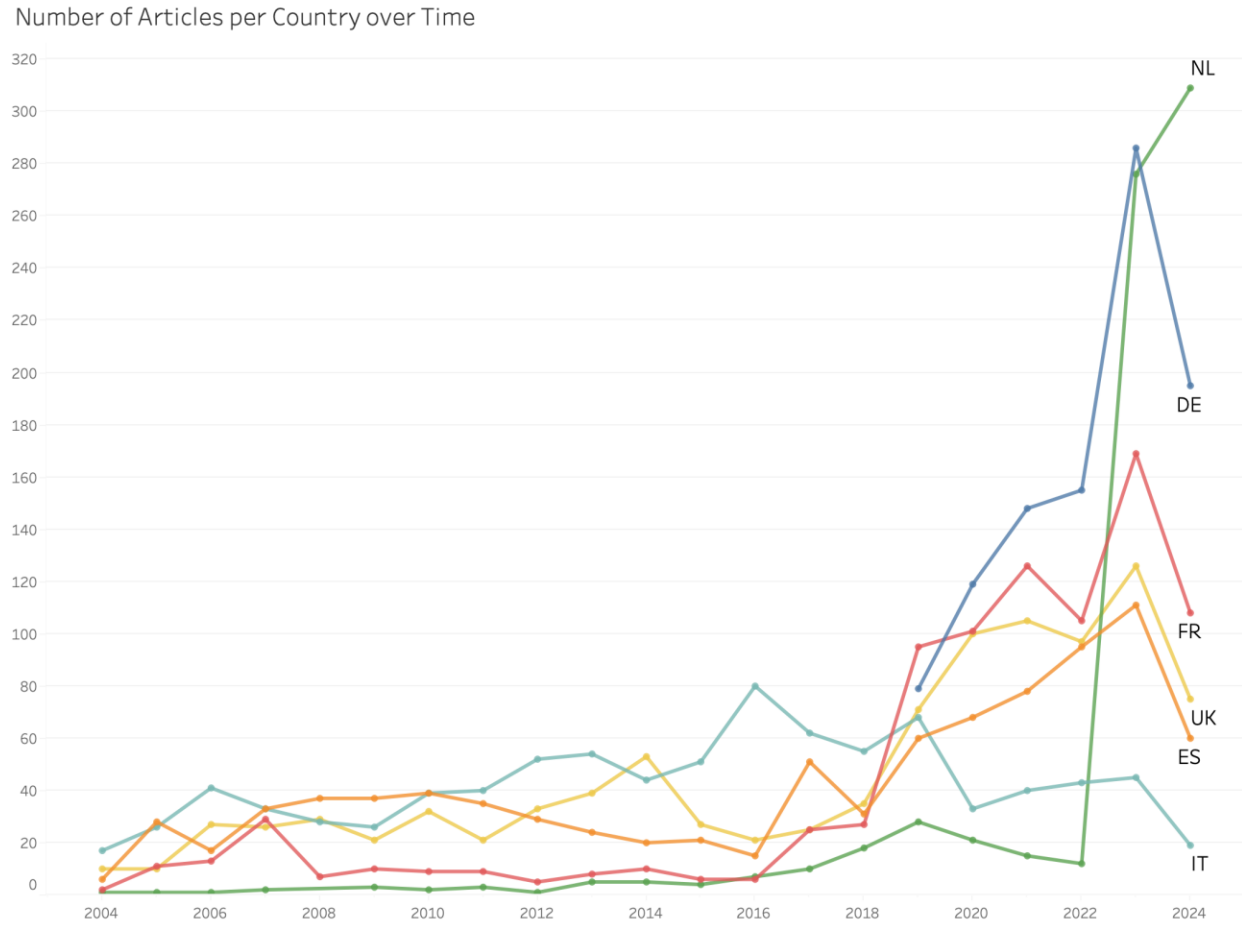
- Extracted articles in RTF format by 100, saved in OneDrive
- Extracted text and article meta data, combined datafiles

Data cleaning

- Shifted columns where mixed up
- Uniform date format
- Translation through Google Translate in GSheets (also possible to do in Python)

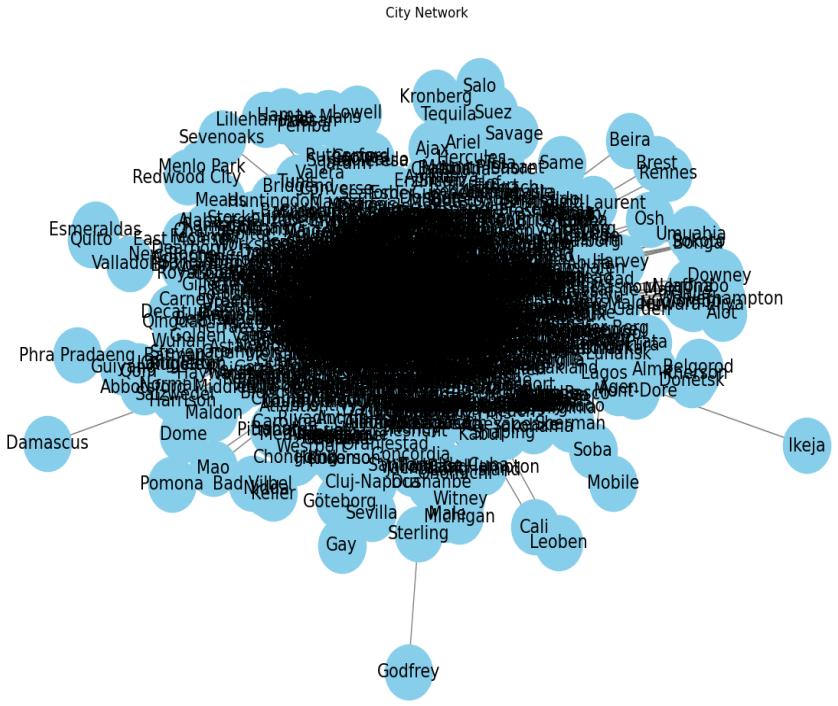
Summary statistics of the data, showing all the mistakes we made

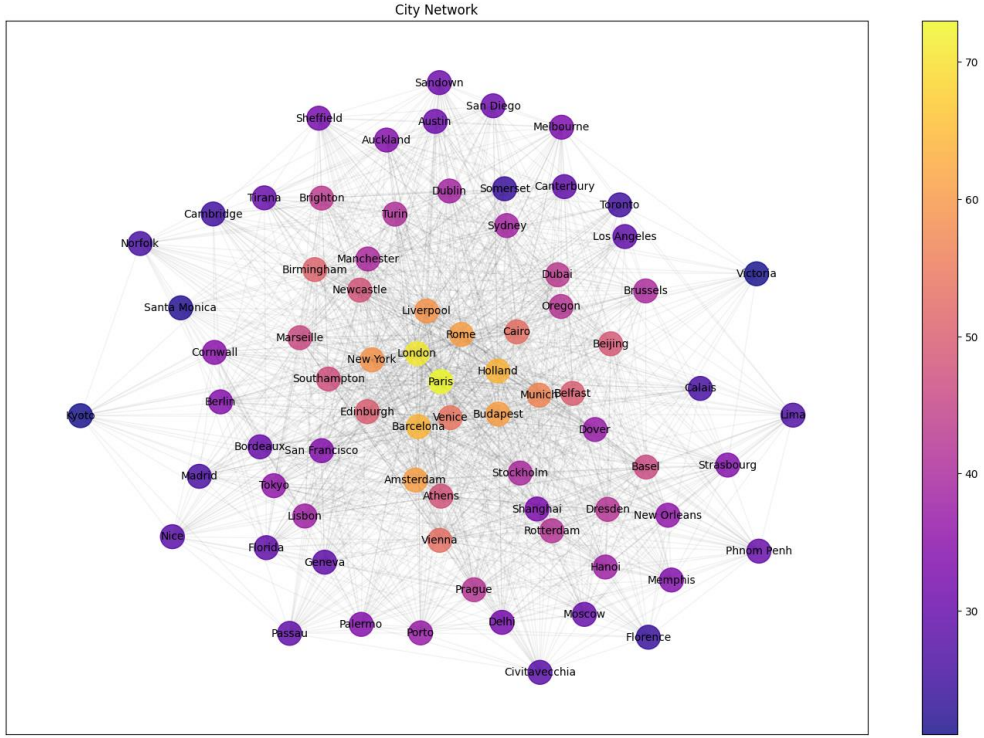




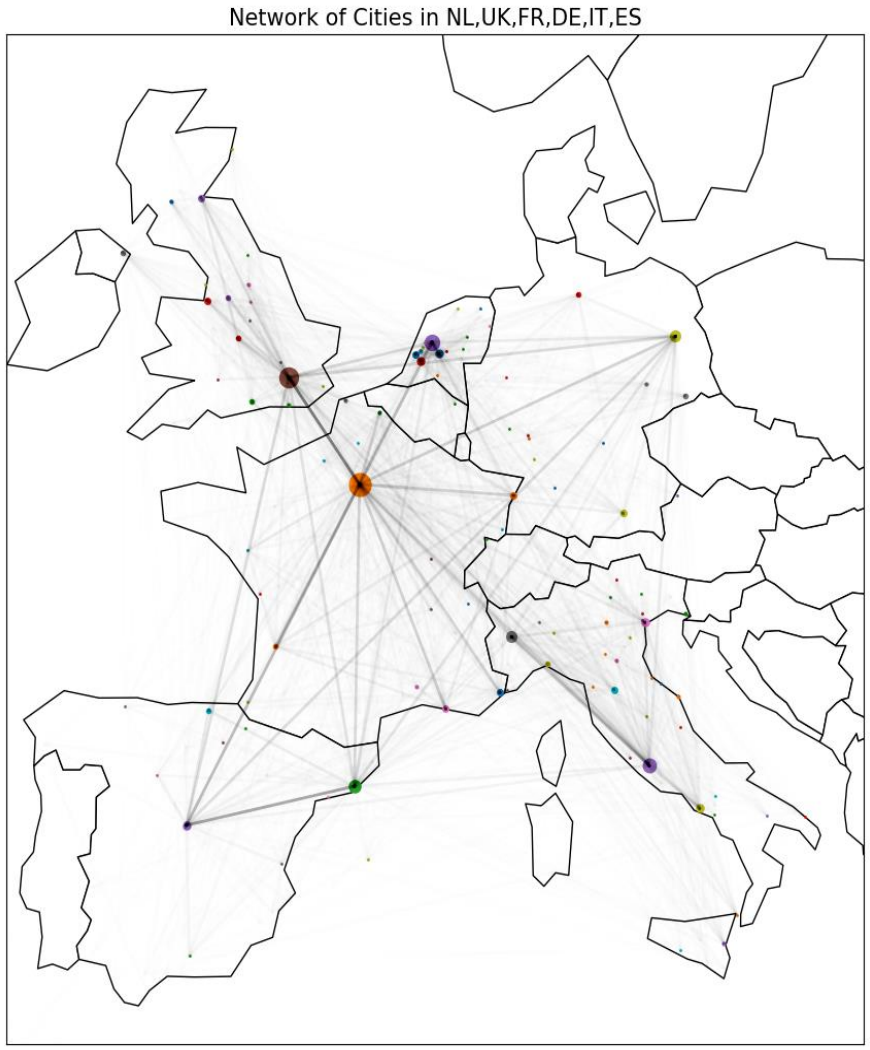
TIME FOR - ANALYSIS!

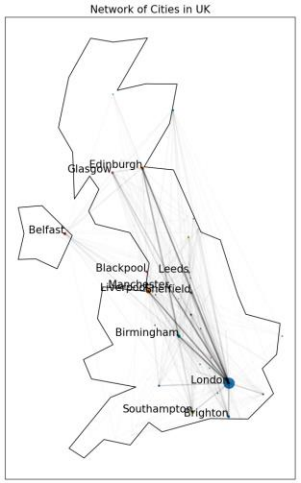
Network Analysis

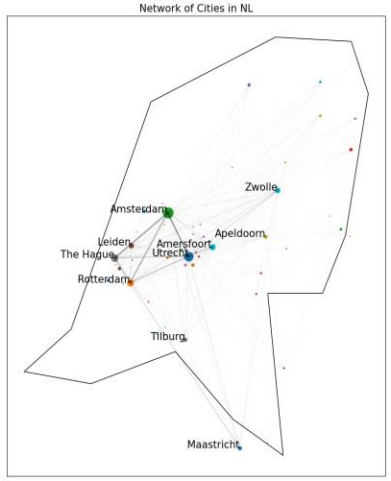


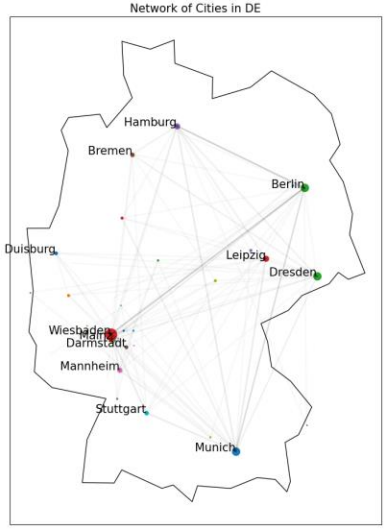


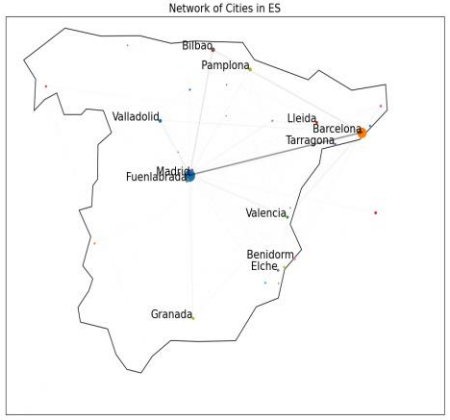
Network of cities – A spatial perspective

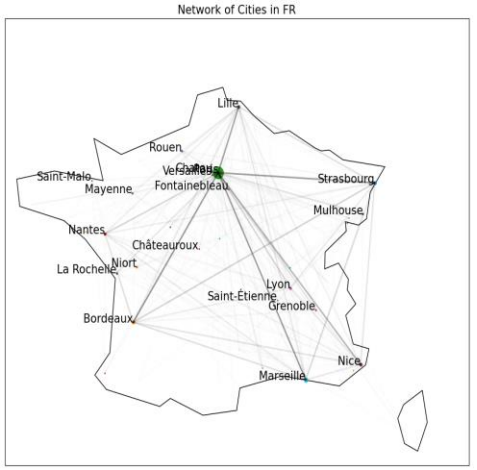


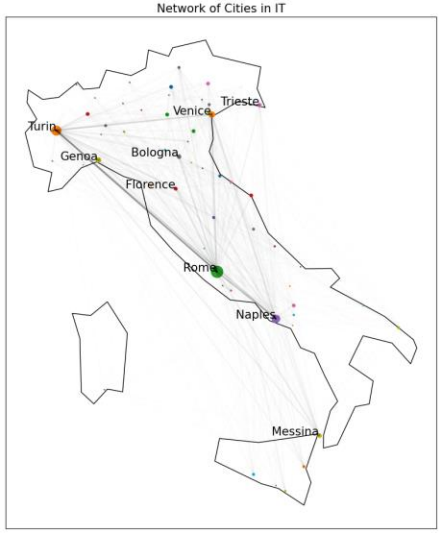










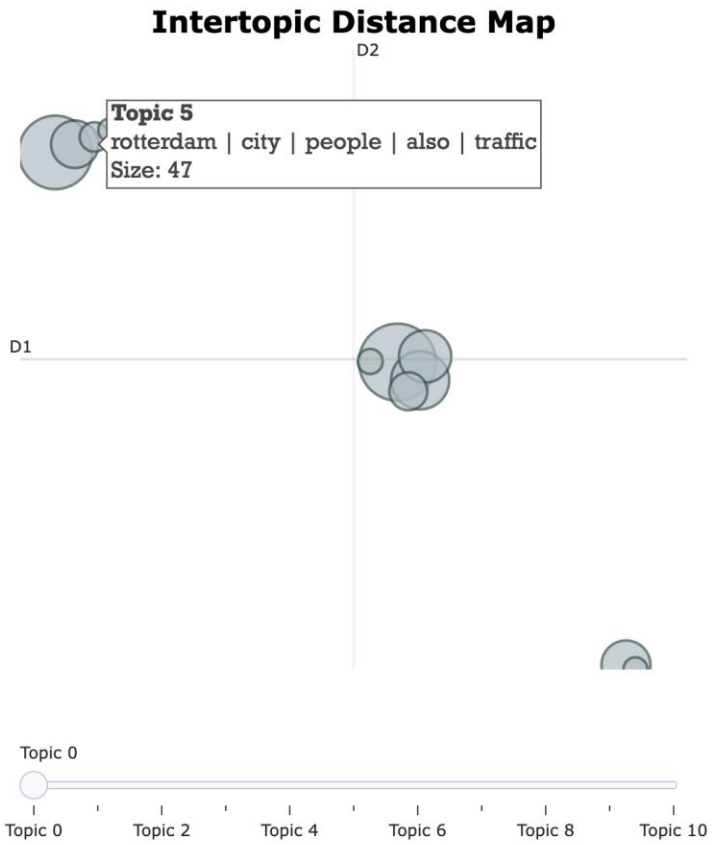


Topics

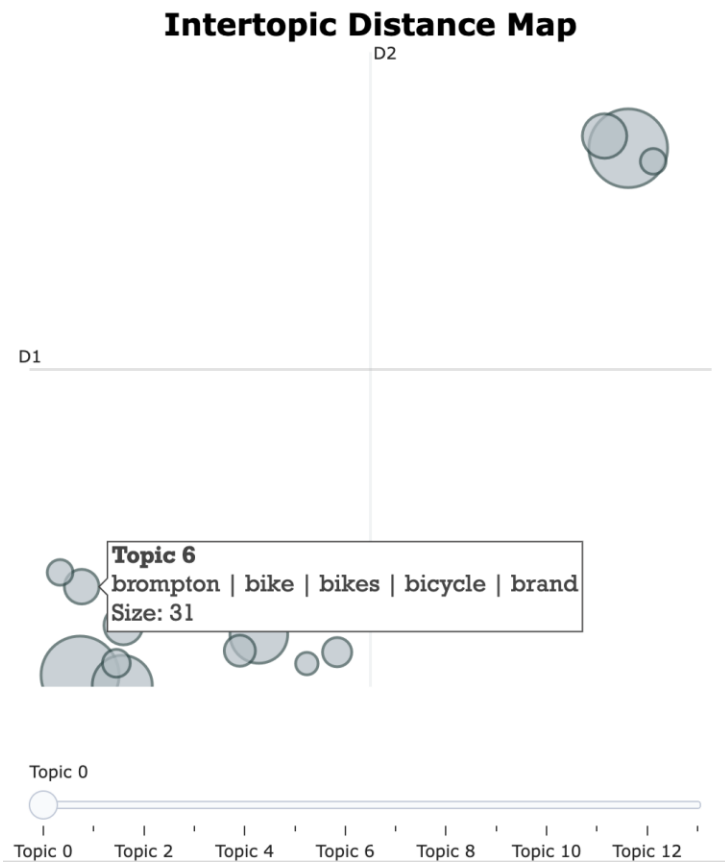
[Topic Modelling]

Topic Modelling

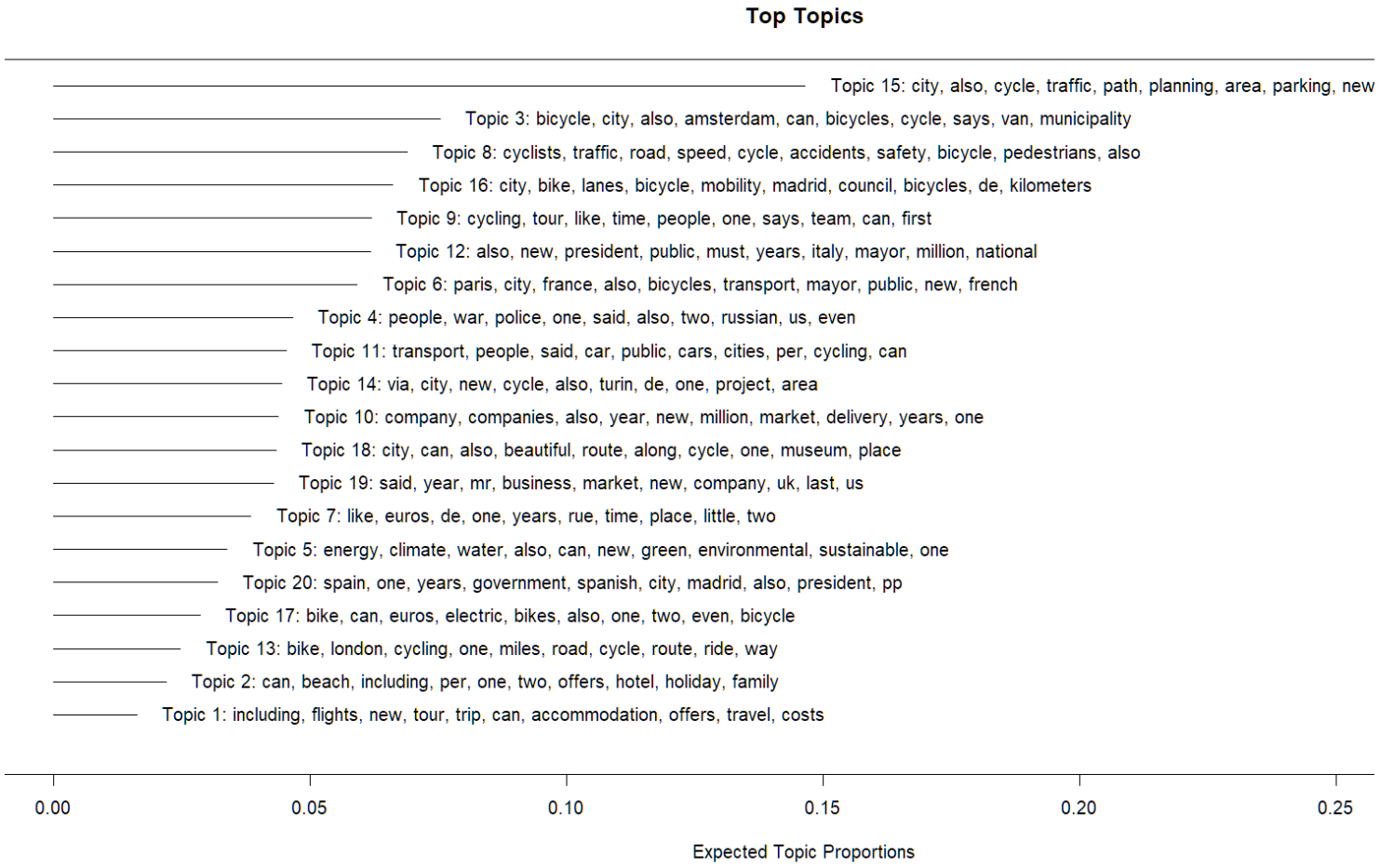
The Netherlands

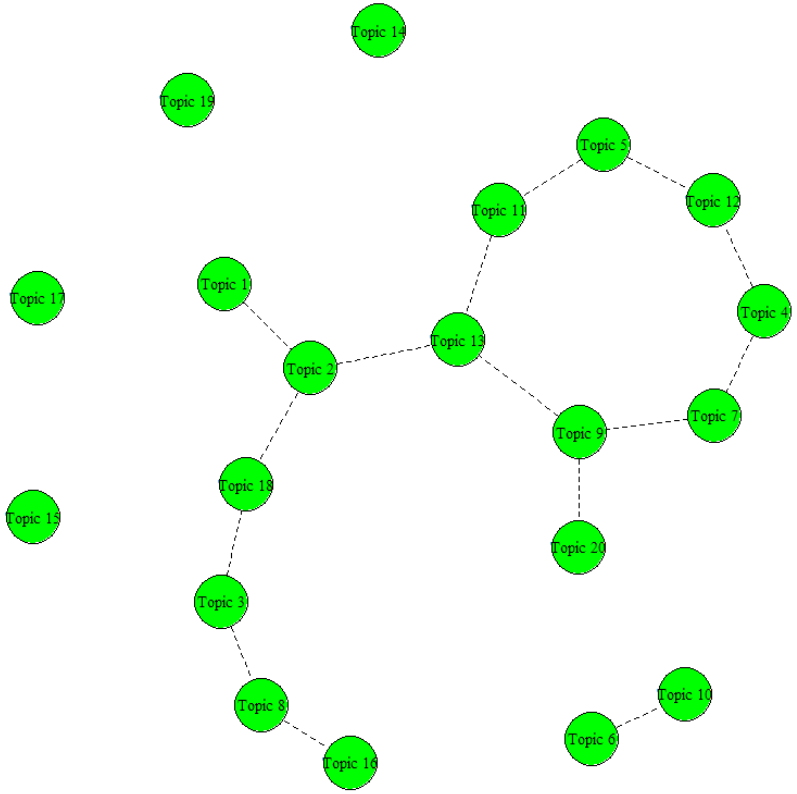


United Kingdom



Structural Topic Modelling (STM)





STM results & interpretation

Topic 1 Top Words:

Highest Prob: including, flights, tour, new, trip, can, accommodation, offers, travel, costs

FREX: lge, cruises, v, flights, departures, voyage, eight-day, cruise, inca, uniworld

Lift: @belvoircastle, @daveyandsky, 1-24, 1-4, 1-5, 1-9, 1-may, 1,000-mile, 1,000-room, 1,000pp

Score: flights, cruises, self-guided, departs, lge, seven-night, eight-day, full-board, sailings, small-group

Topic 2 Top Words:

Highest Prob: can, beach, including, one, per, two, offers, hotel, holiday, family

FREX: t.e, tfe, beach, doubles, sleeps, hideaway, beaches, cottages, hatta, self-catering

Lift: 1h, 24c, 25-minute, 26c, agritourism, airfares, ales, ammonite, anastasia, andros

Score: beach, t.e, self-catering, three-night, cottages, seven-night, tfe, beaches, self-guided, departs

Topic 3 Top Words:

Highest Prob: bicycle, city, also, can, amsterdam, cycle, bicycles, says, cycling, van

FREX: vvd, utrecht, d66, ij, hooijdonk, groenlinks, pvda, amsterdam, rotterdam, hague

Lift: bingöl, bovag, djafarpur, eijck, erkelens, jeroen, langeveld, leidseplein, norg, postma

Score: utrecht, amsterdam, municipality, rotterdam, hague, councilor, bicycle, van, pdf, vvd

Topic 4 Top Words:

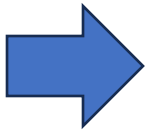
Highest Prob: people, police, war, one, said, also, two, russian, city, even

FREX: putin, gaza, hamas, kyiv, terrorists, palestinian, ukrainian, israel, israeli, weapons

Lift: gaza, @editoreruggeri, 007s, 1.30am, 1.3trillion, 1.7b, 10.20pm, 12-year, 130mph, 133million

Score: hamas, gaza, putin, kyiv, ukrainian, israeli, israel, russian, soldiers, civilians

…….



- Road safety/accidents/ injuries/deaths
- Public transport/mobility services/congestion
- Energy/climate/environment
- Electric bicycles

STM results & interpretation

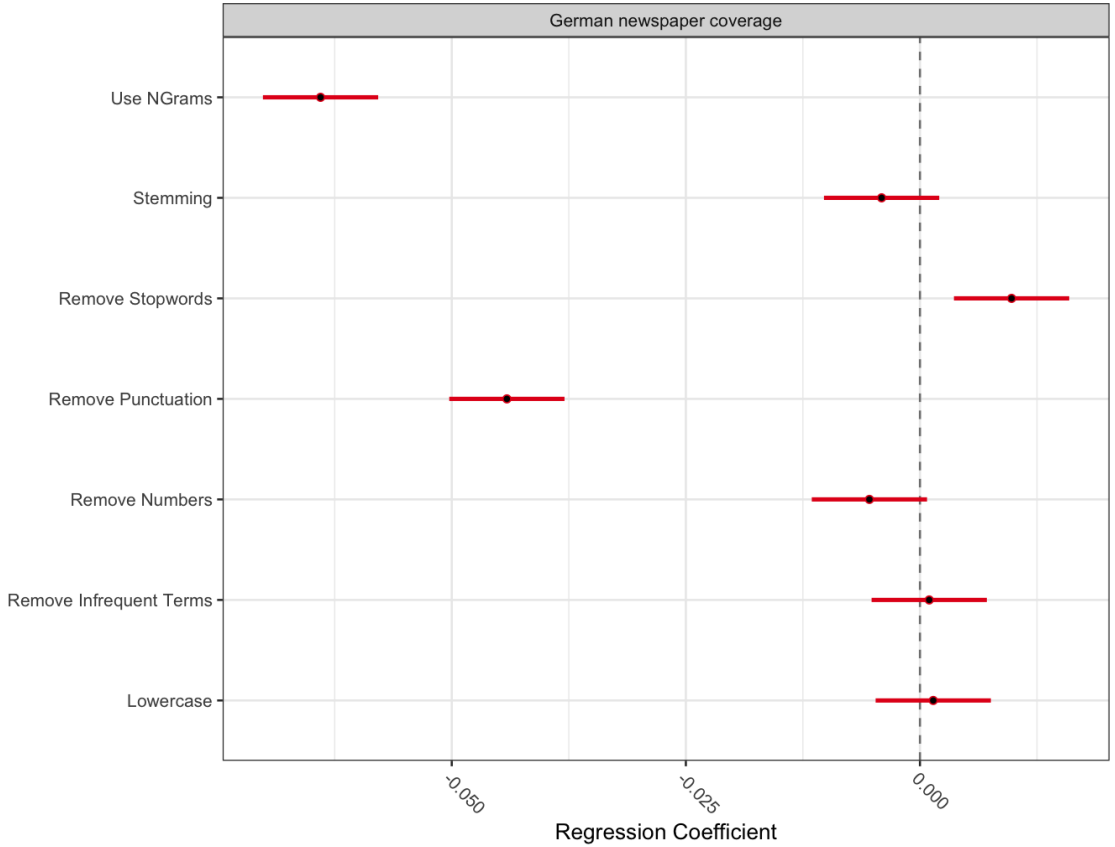
More likely to mention in right-leaning news outlets

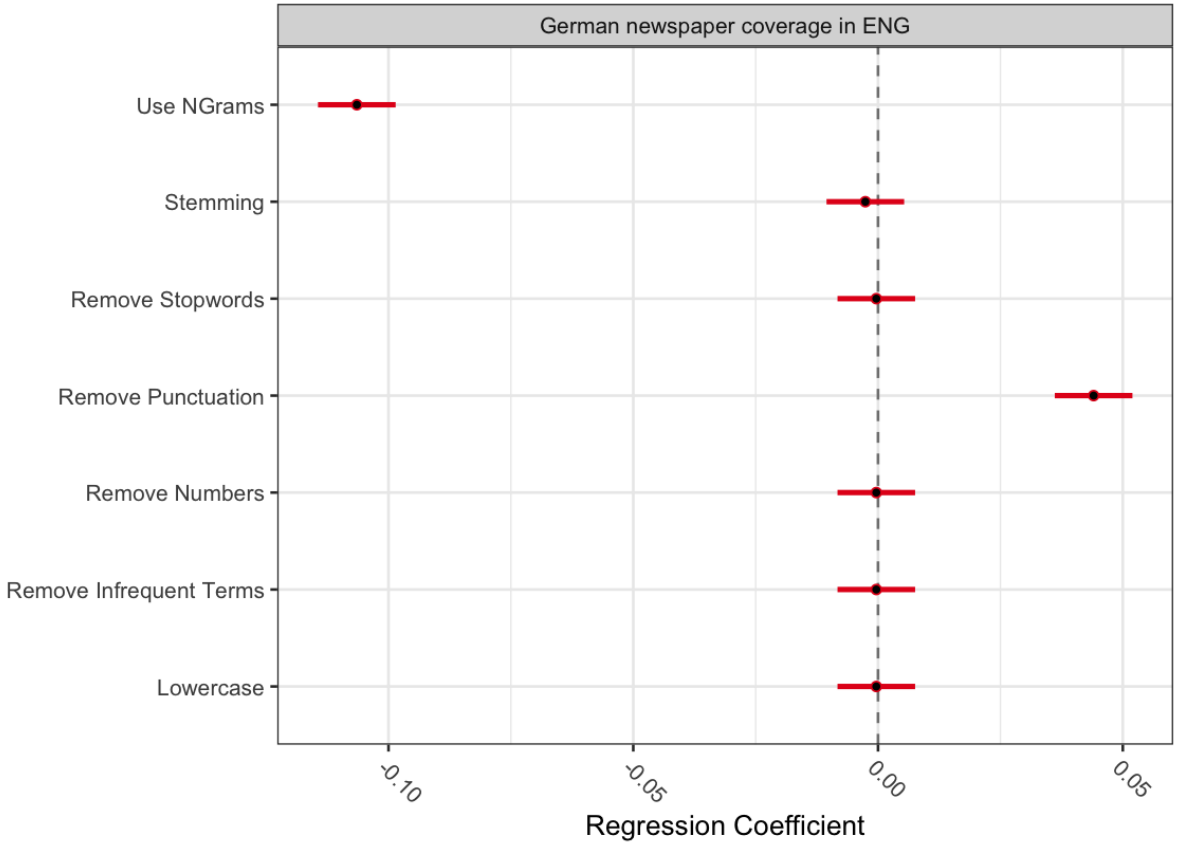
```
Topic 11: Public transport/mobility options/congestion
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.028292   0.004160   6.801 1.16e-11 ***
political_orientationRight 0.026132   0.004913   5.319 1.08e-07 ***
---

Topic 17: Electric bikes
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.029027   0.003975   7.302 3.25e-13 ***
political_orientationRight 0.001728   0.004685   0.369    0.712
---
```

Less likely to mention in right-leaning news outlets

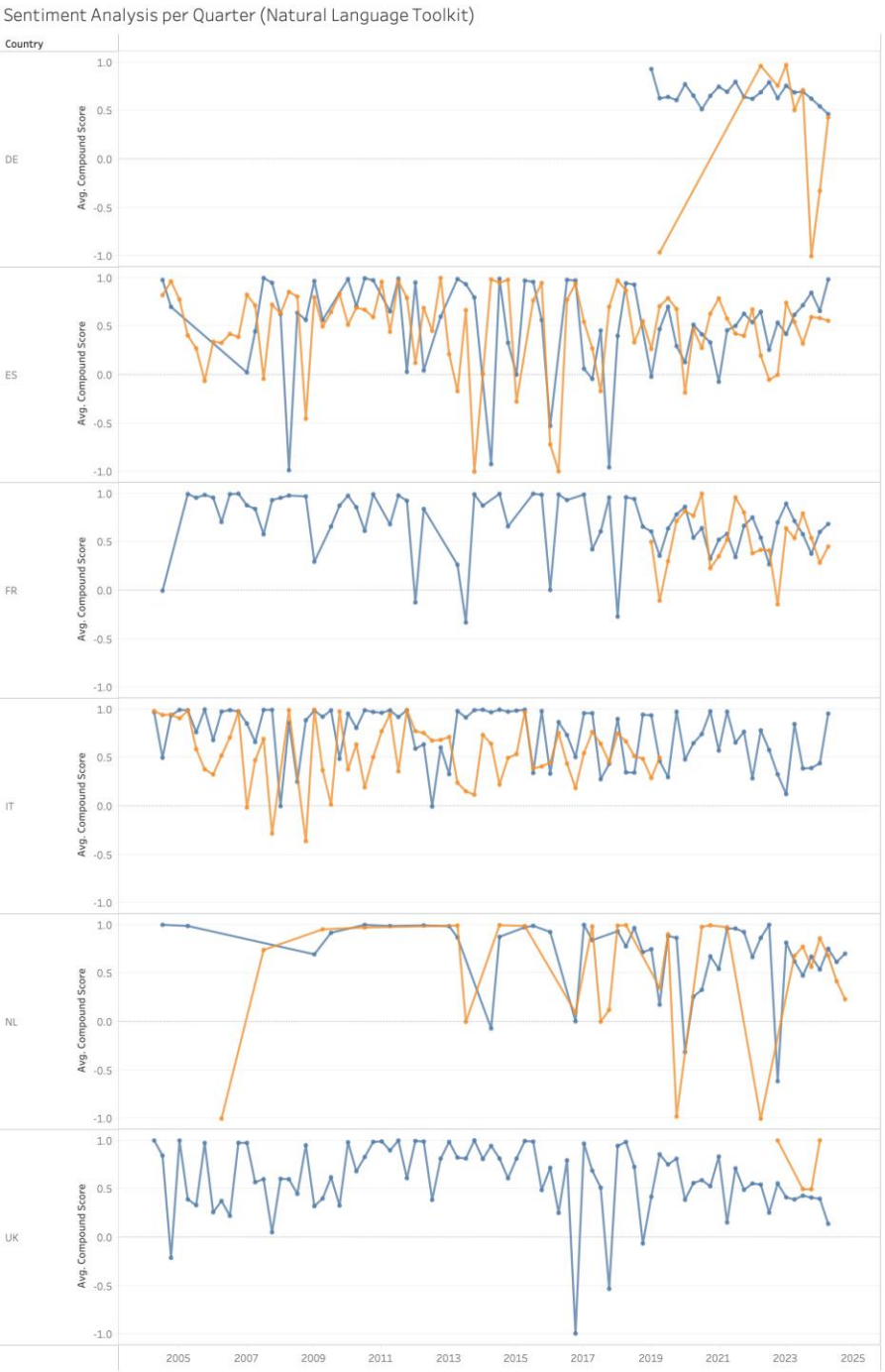
```
Topic 8: Road safety/traffic accidents/injuries/deaths
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.081848   0.004779  17.127  < 2e-16 ***
political_orientationRight -0.019857   0.005469  -3.631 0.000285 ***
---

Topic 5: Energy/climate/environment
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.047186   0.004940   9.551  < 2e-16 ***
political_orientationRight -0.015367   0.005489  -2.799  0.00514 **
---
```

PreText: preprocessing OG and translated text





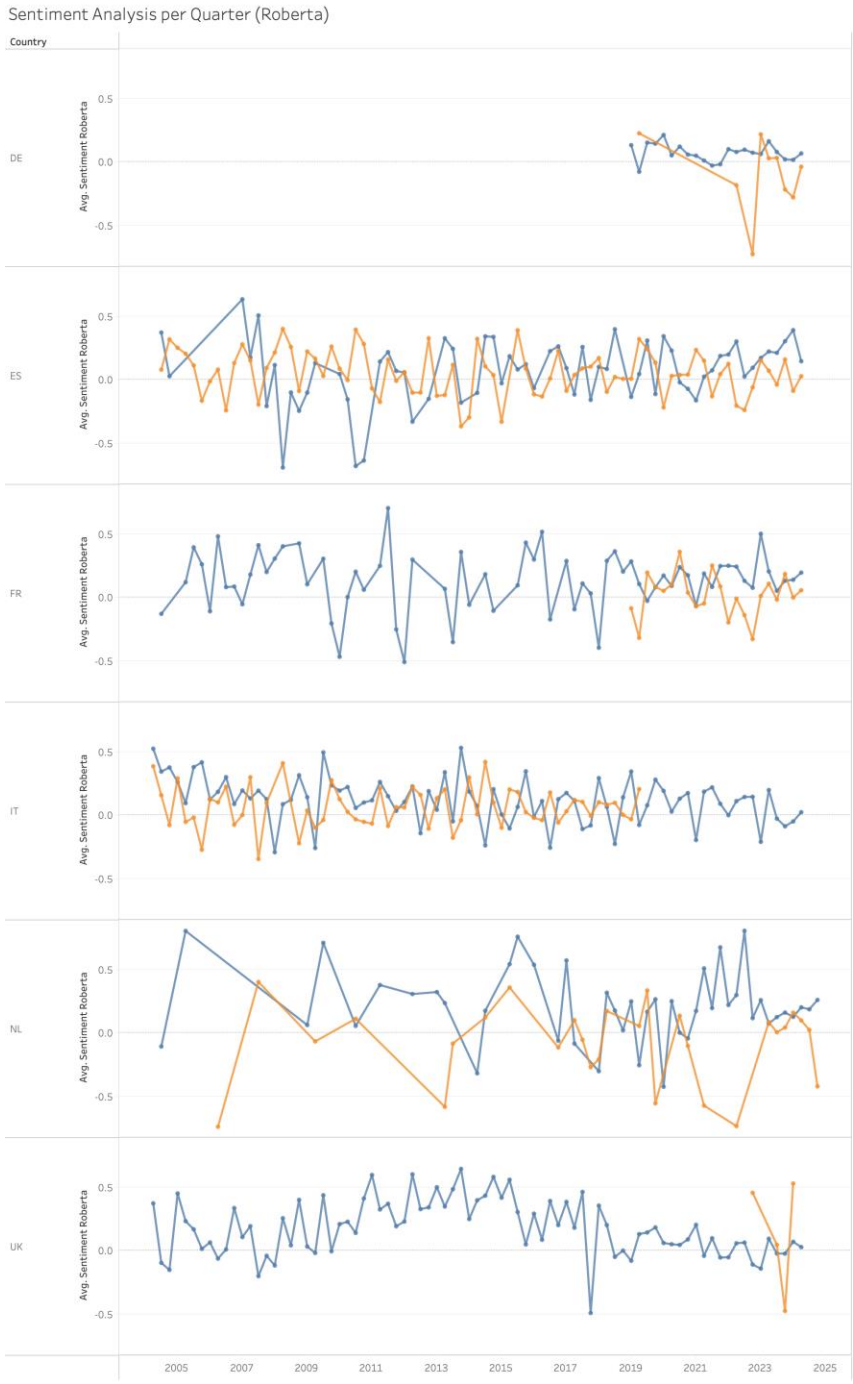
http://www.mjdenny.com/getting_started_with_preText.html



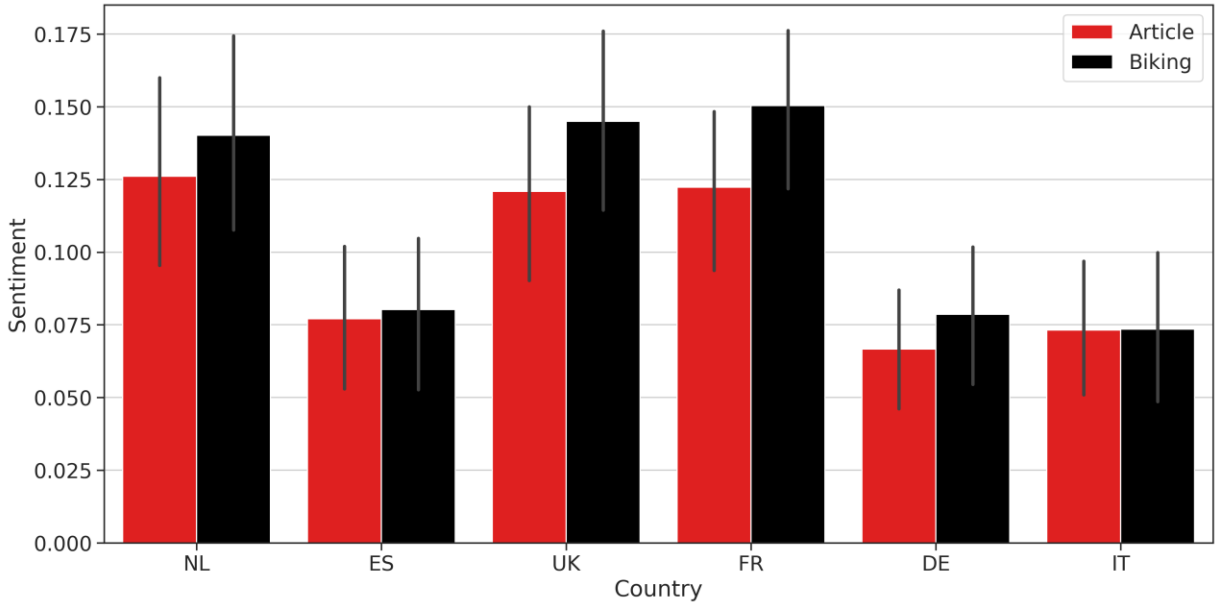
nltk

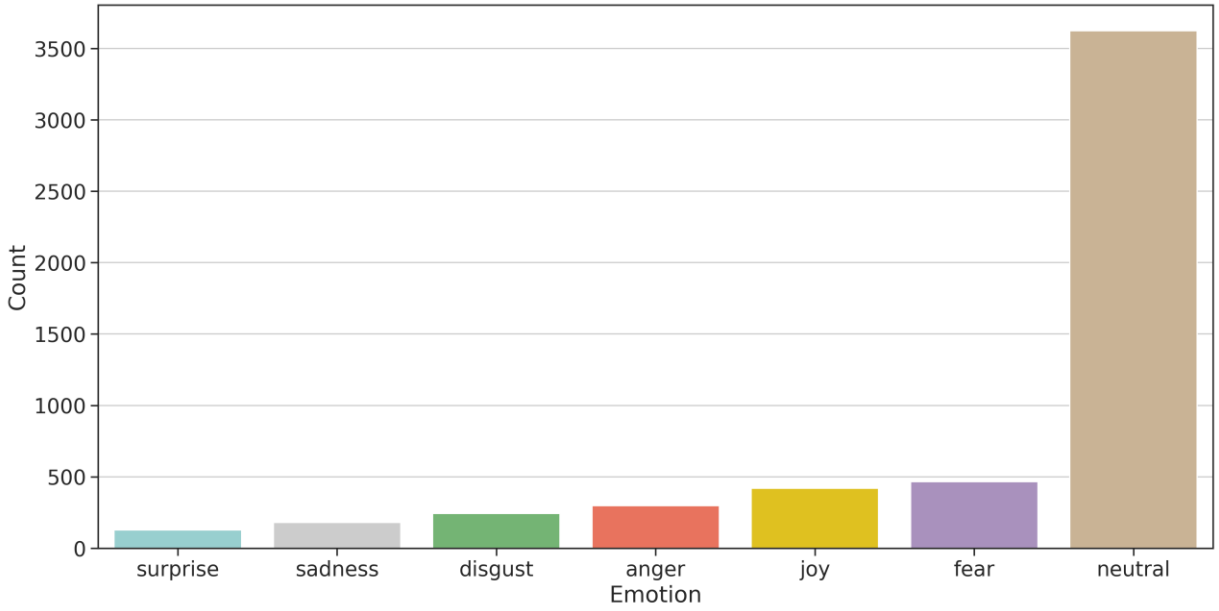
Sentiment Analysis

Roberta



Context-specific sentiment & Discrete emotions





Issues: real-life data & collaboration are messy

IMPORT

- Download limits from LexisNexis – 1000 articles/account (and we *accidentally* did it on "relevance")
- Not all newspapers available – had to take less relevant newspapers
- Search terms – "cycle" is also an "electoral cycle", "life cycle", etc.
- Searched both newspapers, sorted by relevance – more from one newspaper than the other
- Same newspaper had multiple different names (Daily Telegraph)

PROCESSING

- Accents on some letters in some languages -> preprocessing is language-specific
- Translation of articles
    - Google sheets solution
    - … but some imperfect translations ("firefighter" -> "fireman")
- Combining datasets – shifted columns (7% incorrect) due to translations and UK data – had to write shotgun code
- Multiple people working on the same dataset, dealing with multiple push requests/overwriting sections of code. Also, no time for factoring the code and **creating a general solution** to come up with consistent scheme.

GENERAL COWORKING

- Our code is a mess. Some efforts in trying to refactor and so on, but as we were working across several languages etc. It was bit tricky - it is very easy to become chaotic.

Conclusions of the project

“A man might befriend a wolf, even break a wolf, but no man could truly *tame* a wolf.”

— George R.R. Martin, A Dance with Dragons

Like

tags: skinchanger, warg, winter-is-coming, wolves

394 likes