Sweden from the Eye of GDELT

Members:

Yaser Kaddoura

Mohammad Mustakim Ur Rahman

Supervisor:

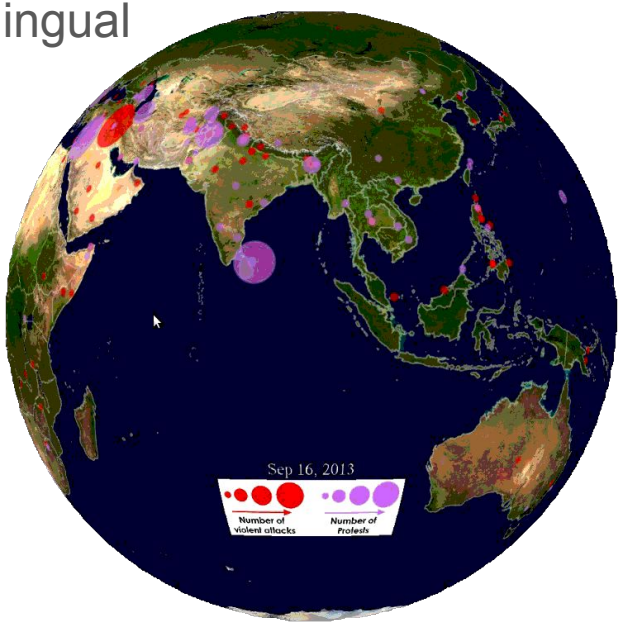
Raazesh Sainudiin

Overview

- Introduction to GDELT
  - GDELT 2.0 Datasets
  - Events CAMEO taxonomy
- Experiments
  - Events CAMEO taxonomy in Sweden
  - Comparison between Sweden and Norway
  - Shootings
  - Reliability

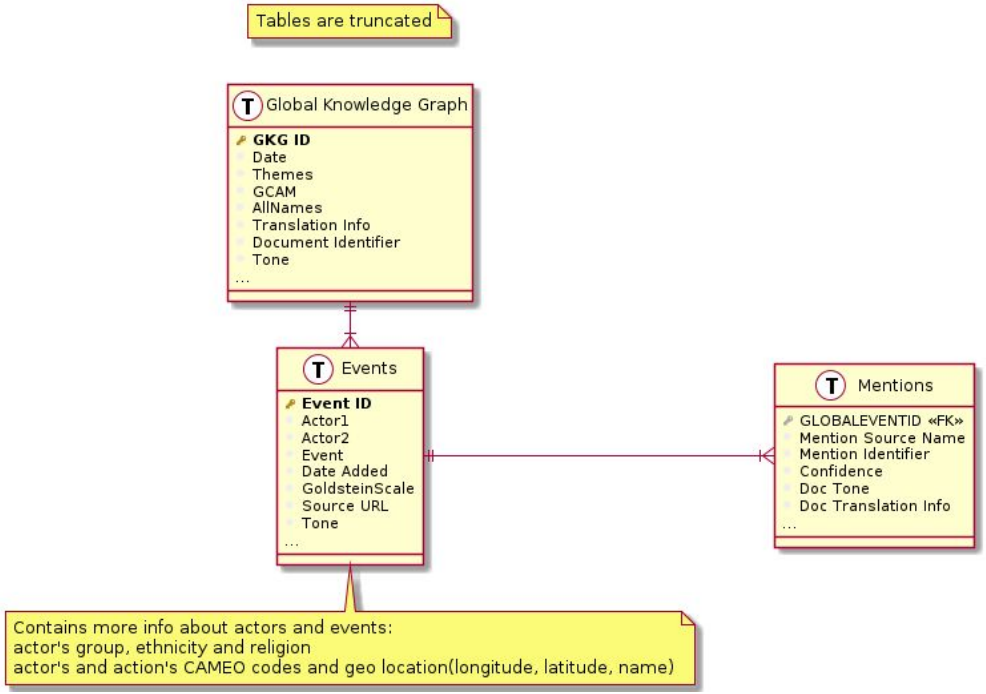
THE GLOBAL DATABASE OF EVENTS, LANGUAGE AND TONE (GDELT)

- Events of the entire world at your fingertips since 1979
- Access over 65 languages through GDELT Translingual
- Sentiment analysis
- More than 2,300 emotions and themes
- An attractive source for researchers



Some datasets

President Reagan has **threatened** further action against the **Soviet Union** in an international television program beamed by satellite to more than 50 countries.



Sentiment Analysis

- Pipeline that uses linguistic tools
- Assesses 2300 dimensions (GCAM)
- Calculates overall tone of text

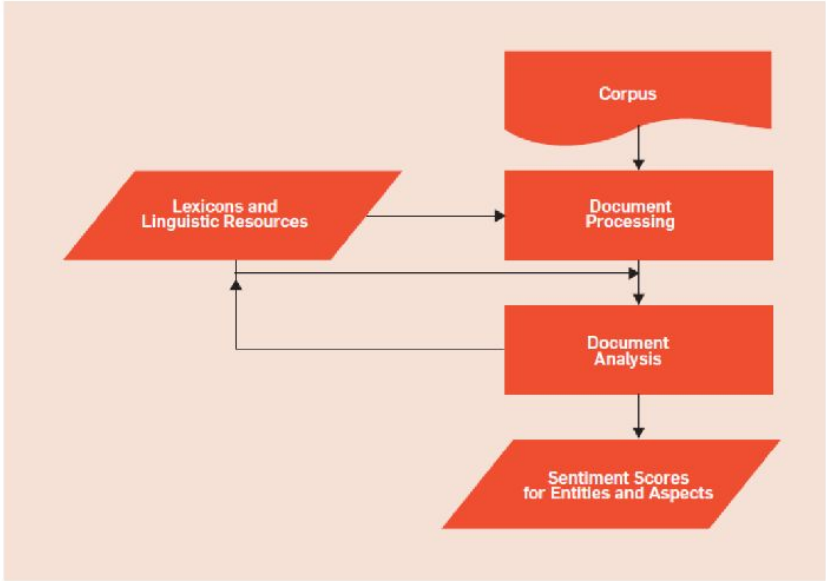
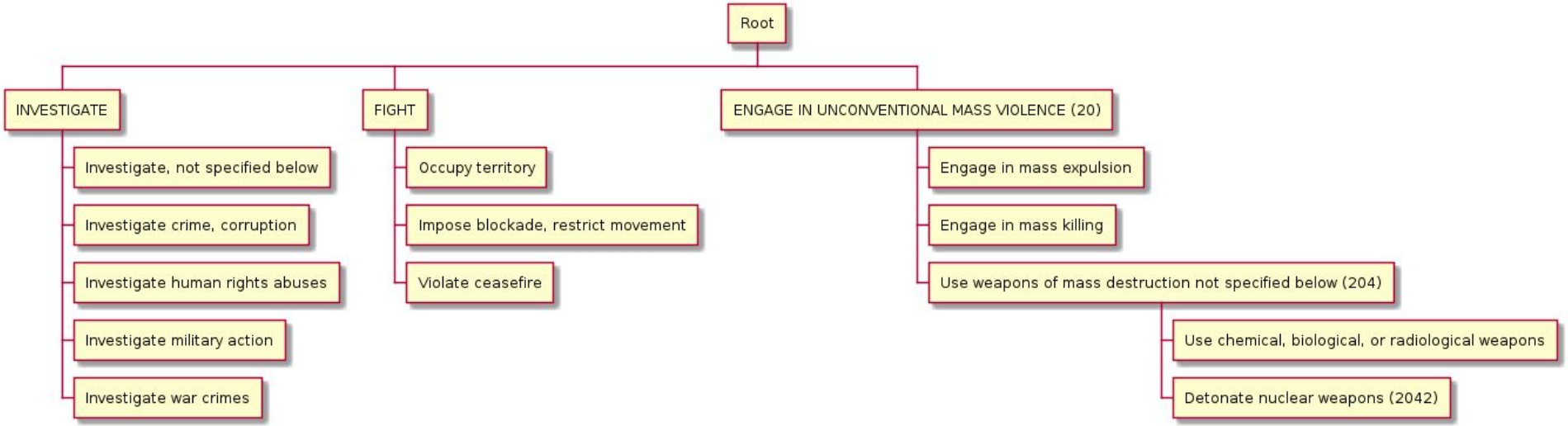


Figure 1. Architecture of a generic sentiment analysis system. Retrieved from Feldman (2013).

Image taken from: https://www.globe-project.eu/the-empirical-use-of-gdelt-big-data-in-academic-research_13809.pdf

Events CAMEO taxonomy

- Represented by at most 4 digits
- Three levels
- Events have more context the lower you go down in the tree



Experiments on GDELT

- Using Apache Spark to handle the massive amount of data in GDELT 2.0
- GDELT Translingual (Non-English documents)
- Data for the 2021 year
- Events inside the country only
- Events that has negative **Goldstein scale**

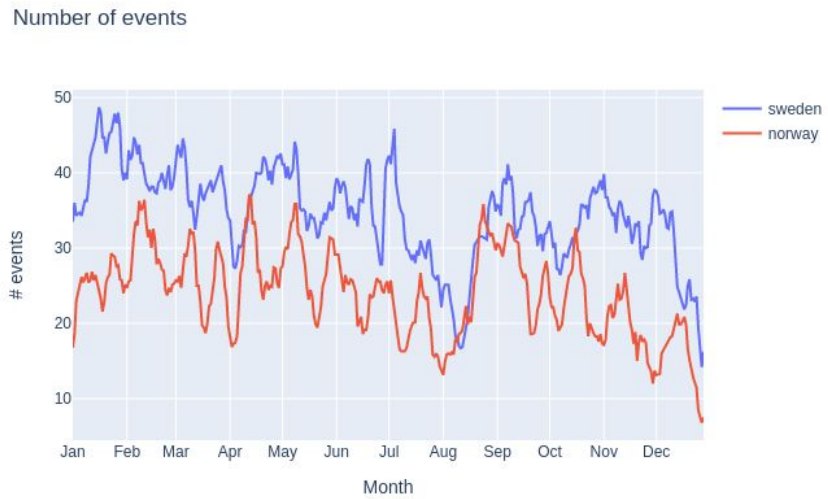
Goldstein scale: A [-10, 10] score that identifying the theoretical potential impact of the event on the stability of the country

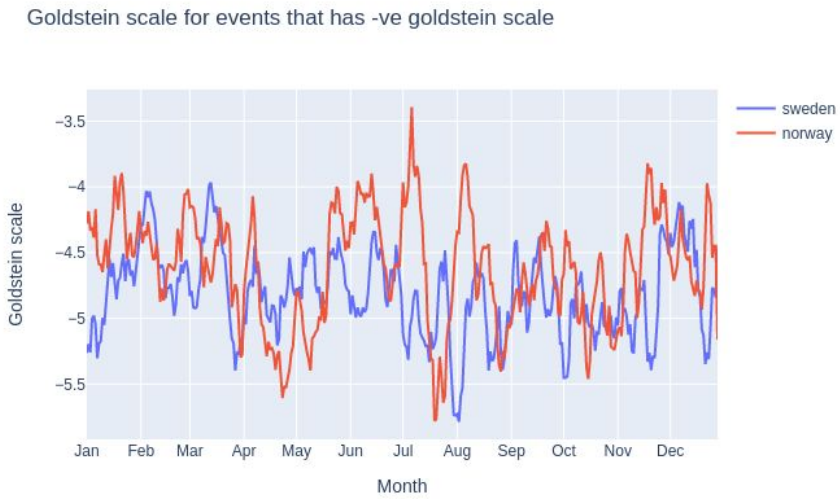
Events CAMEO taxonomy in Sweden (Treemap)

- 2021 Swedish media
- Color represents the tone of the text in the articles
- Popular events: INVESTIGATE (3533), FIGHT (3251), DISAPPROVE (3014)

Comparison between Sweden and Norway (Time series)

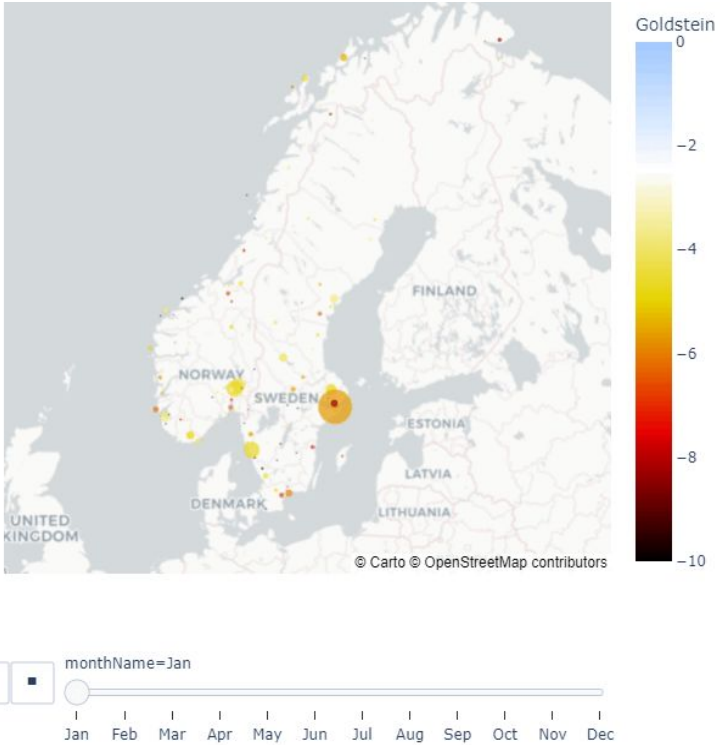
- Norwegian and Swedish media only
- Sweden media covers more events than Norway





Comparison between Sweden and Norway (Geo map)

- Bubble size represents number of events
- Color represents Goldstein scale
- Cities that has the most events in 2021:
  1. Stockholm (2808)
  2. Goteborg (906)
  3. Oslo (712)



Shootings in Sweden

Extracting events that’s related to shootings in Sweden.

Pipeline has the following steps:

1. Connect to the analytics/AI-ready GDELT delta lake house
2. Fetch URLs for relevant articles presented in GDELT
3. Scrape the text from articles
4. Translate articles to English
5. Perform NLP techniques to extract terms
6. Check if terms related to the shooting is present (e.g. shot, gunshot)

Shootings in Sweden (Time series)

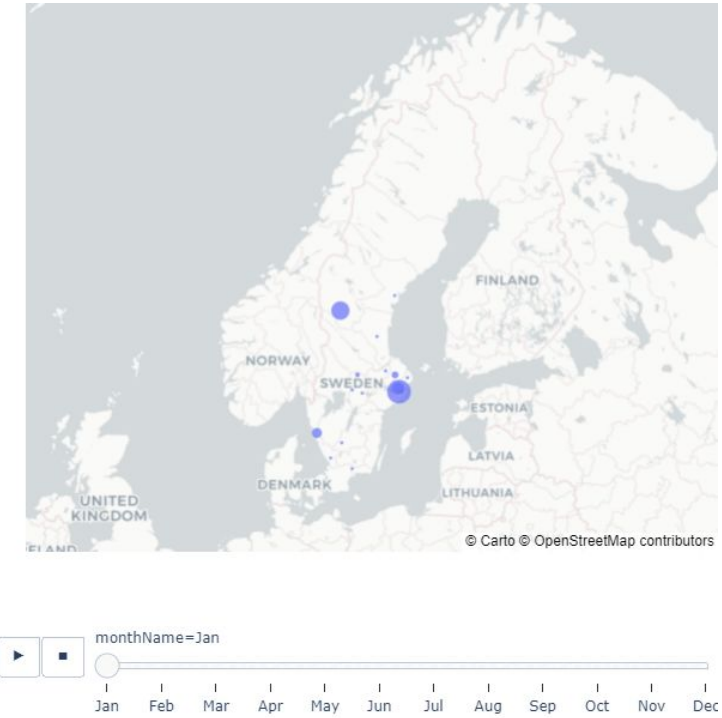
Months with most shooting activities:

1. July (241 events with 378 mentions)
2. October (205 events with 349 mentions)

Shootings in Sweden (Geomap)

Cities that has the most shooting events in 2021:

1. Stockholm (568)
2. Harjedale (425)
3. Goteborg (194)



Reliability

- GDELT’s erroneous processing
  - Different keys for same events
  - Misidentified events (COVID shots -> shootings)
- Naive pipeline for extracting shootings
- More sophisticated NLP pipelines can be used to reduce the error

Conclusion

- Most likely, you won’t find what you need from GDELT
- Regardless, it’s a resource worth investigating
- Using NLP pipelines on top of GDELT
- Without any clean data to work with, GDELT could be a starting point