Submitted By: Shankar Acharya (sa225xp@student.lnu.se)

About Dataset

The data set I am interested in is called the **Ukraine Conflict Twitter Dataset**, it contains tweets monitoring the current and ongoing Ukraine-Russia conflict since February 23, 2022.

Size of dataset: 47.5 GB(overall)
Number of rows: ~1.2M
Overview of data: Contains actual tweets regarding the conflict with multilingual text and hashtags, and mentions

DataSet Description

| S.N	Field	Type	Description
1	user_id	integer	User ID of the Twitter internal system
2	username	string	Twitter visible username
3	acctdesc	string	Tweet posted by User ID(content, hashtag, mentions)
4	location	string	Name of the place the user is from
5	following	number	Number of followers the User ID has
6	totaltweets	number	Number of tweets posted by User ID till the data was tracked
7	usercreatets	number	Timestamp when User ID was created in twitter

User Interest

Here is a list of insights, users would want to get from this dataset
- What are the pro-NATO countries & pro-Russia nations?
- Has any nation switched from being pro-NATO to pro-Russia over time?
- How has the sentiment increased/decreased over time?
- Has Trump's re-election changed the sentiment?

Conceptual Design

The conceptual design of the visualization would look like
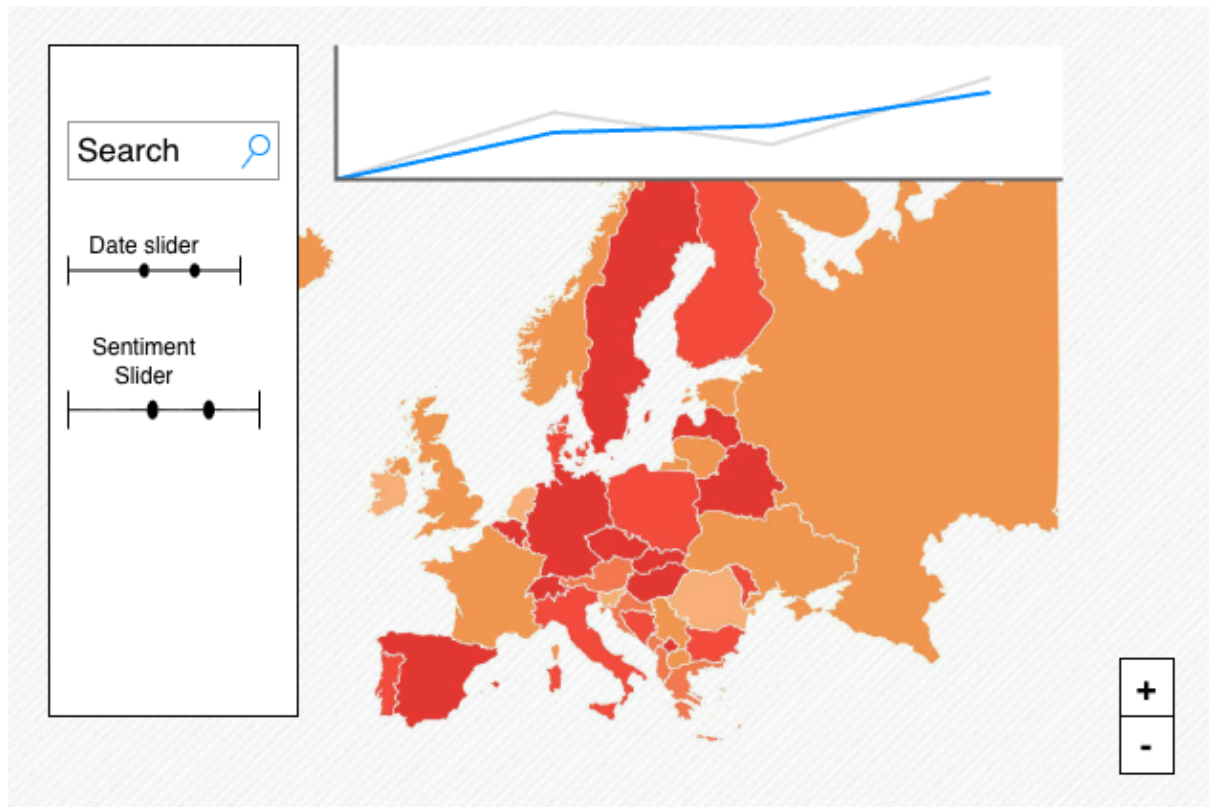
We will have multiple coordinated views. Users will have a shared filter where they can filter regions by region/country, slide between different dates, and filter sentiment. One visualization shows the sentiments on the map while another shows trends over applied filter. We will use color encoding to represent if a nation is pro-NATO or pro-Russia. Darker color would mean a higher weight.

User Task/ Action

S.N	Task	Action
1	Find a change in sentiment across time	Slide the date slider and change the start date and end date
2	Filter regions by sentiment	Slide the sentiment slider
3	Select the sentiment in a specific country	- Search the selected Country in the Search bar. - Pan and zoom over the map
4	Get total pro-NATO, pro-Russia, or neutral tweets in a specific country	- Hover over a specific country, and the tooltip displays the count

How can the visualization answer user interest?

- What are the pro-NATO countries & pro-Russia nations?
  The user sees different colors applied to distinguish which nation is pro-NATO and pro-Russia, and a varying gradient to see which country is more influenced than to other.

- Has any nation switched from being pro-NATO to pro-Russia over time?
  The user uses the date slider and observes the trend graph over time. Optionally observes a change in color in the map.

- How has the sentiment increased/decreased over time?
  The user uses the date slider and observes the trend graph over time for the selected country.

- Has Trump's re-election changed the sentiment?
  The trend graph will have a mark line pointing to the time when Trump was re-elected.

Data Processing and System Overview

The designed system would need the following data preprocessing:
- Filter only specific tweets that are in English, parsable, and meaningful.
- Clean tweet data by trimming hashtags, mentions, and links added in the tweet.
- Feed the tweet to the LLM model to understand the sentiment with score.

From the above data processing, we would have the right data needed for visualization. Then the output data would be populated and used in the app for visualization. The implementation details:

- Backend: Python FastAPI; Postgres
- Frontend: TypeScript + React (Vite); D3.js for trend graphs; Mapbox GL for maps;
- NLP: Hugging Face Transformers + BERTopic; sentiment (e.g., XLM-RoBERTa variants);
- Deployment: Docker Compose locally;