Authors:
Mikko Laitinen
Paula Rautionaho
Masoud Fatemi
Mehrdad Salimi
University of Eastern Finland
Contact: {first.lastname@uef.fi}

FIN-CLARIAH 2024/25

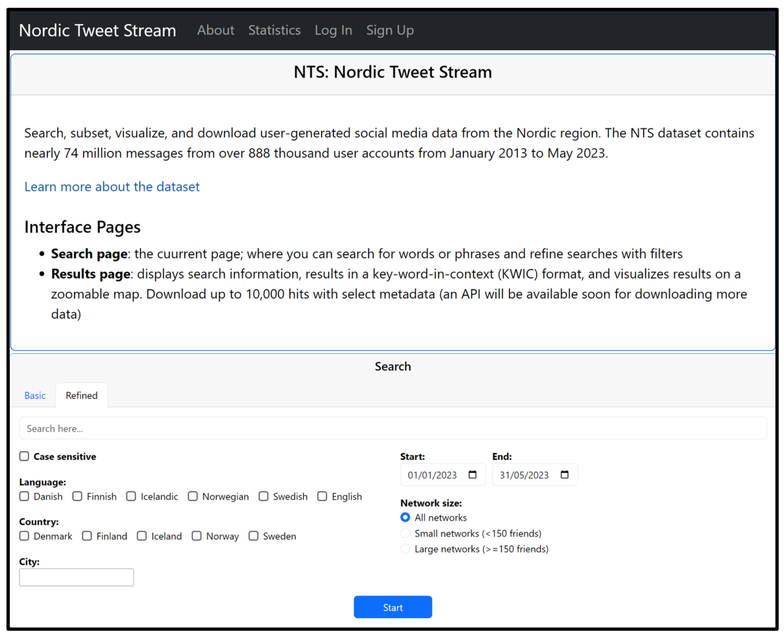
D:4.1.2 & D:4.1.3 Representative benchmark data of social media and digital tools for network analysis

What and why?

According to recent estimates, the majority of people globally use social media. Yet, representative datasets of these platforms are rare. Our objective is to create benchmark collections across various social media platforms and design new digital tools that could provide novel insights into old questions in social network analysis in SSH.

What is NTS?

Nordic Tweet Stream (NTS) is an online tool that enables searching, subsetting, visualizing, and downloading user-generated social media data from the Nordic region (D:4.1.2). It features multilingual content and is designed for use by researchers from diverse fields including sociolinguistics, dialectology, social sciences, and cultural studies.



Search page of the NTS web application

NTS data

NTS is a large collection of geolocated tweets and related metadata from the Nordic area, serving as a digital corpus. It emphasizes the necessity of independent storage of large social media datasets to ensure they remain accessible for research. These data were gathered using the now-defunct academic Twitter API and are now stored at CSC.

NTS statistics

Texts and metadata from January 2013 to May 2023, including ~800 million words from over 888,098 users in 73 languages. The largest languages are Swedish (c. 31 %), English (c. 26 %), and Finnish (c. 13 %).

Country	Tokens	Tweets	Accounts	Locations	Languages
Denmark	109,512,792	9,223,463	232,917	5,189	71
Finland	145,175,345	14,843,735	180,600	18,473	68
Iceland	20,598,240	2,012,410	87,711	1,784	62
Norway	122,304,865	10,939,565	201,956	5,568	72
Sweden	401,576,803	36,751,873	369,380	9,959	72
Total	**799,168,045**	**73,771,046**	**888,098**	**40,973**	**73**

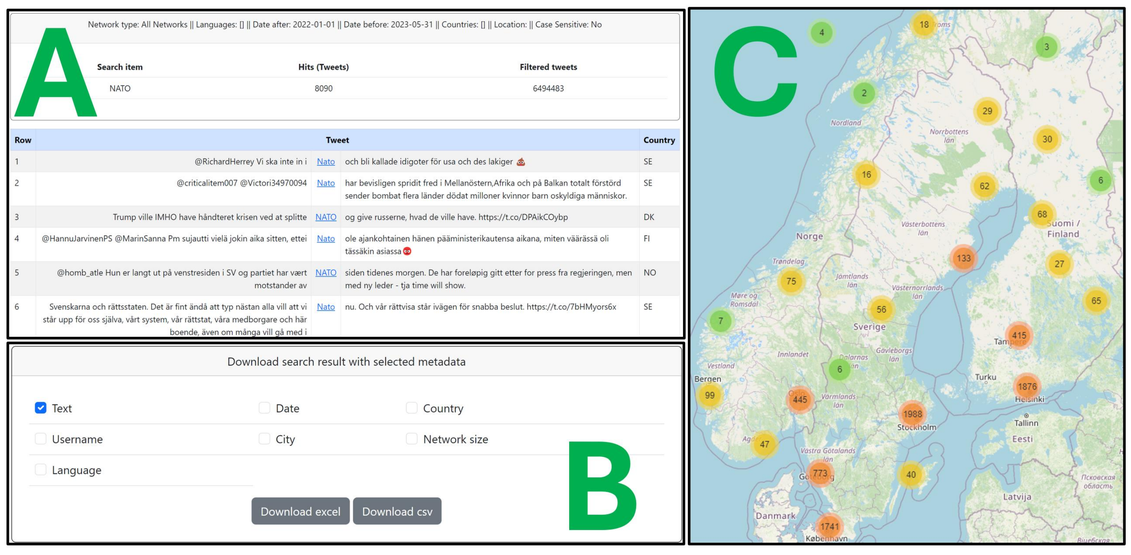
Basic NTS statistics: Jan 2013-May 2023

NTS search interface features

A: Output of the search and its results in KWIC view

B: Up to 10,000 hits available for downloading in two formats with selected metadata

C: Distribution of the search results plotted into an instructive map



Digital tools for social network analysis

This WP focuses on creating digital tools that enhance social network analysis in data-intensive sociolinguistics. It results in four extremely large-scale social media datasets for research purposes, each enriched with network properties. The datasets consists of the Nordic Tweet Network (NTN) and comparable datasets (a.k.a sister datasets) from Australia (AU), the United Kingdom (UK), and the United States (US). They contain nearly 11 billion words from 759,495 people in 19,345 networks. Below are a few sample networks from these datasets.

Nordic Tweet Network (NTN) **Bristol** **New Orleans**

D:4.1.2 TASKS for 2024

We have transferred most of the Nordic Tweet Stream (NTS) data from Rahti 1 to cPouta. We are working on improving the tool's performance and adding rest of the data to the database. We are also in the final stages of developing an API that supports retrieving results, allowing downloads of up to 100,000 hits for more advanced users. Additionally, we are implementing new functionalities specifically for advanced text analysis to broaden the tool's analytical capabilities.

D:4.1.3 TASKS for 2024-25

We have developed an algorithmic solution to quantify network properties of the four datasets. We explore the optimal programming stack, data structures, and platforms to begin developing a web application to make the data searchable. We simultaneously enrich these network datasets with machine learning solutions. These advancements are specifically designed to predict the age range of users and classify them into different user categories.

UNIVERSITY OF EASTERN FINLAND

DARIAH-FI

KIELIPANKKI The Language Bank of Finland

FIN-CLARIN