FIN-CLARIAH

Created by:
Prof. Mikko Laitinen
School of Humanities
University of Eastern Finland
Contact: mikko.laitinen@uef.fi
Mr. Masoud Fatemi
School of Computing
University of Eastern Finland
Mr. Mehrdad Salimi
School of Computing
University of Eastern Finland
Dr. Paula Rautionaho
School of Humanities
University of Eastern Finland
Dr. Antti-Jussi Kouvo
Department of Social Sciences
University of Eastern Finland

WP 4.4 SOCIAL MEDIA NOISE (and how to tackle it?)

BACKGROUND

Various social network applications (Discord, Facebook, Reddit and Twitter, etc.) have turned the web into a user-generated repository of information in a number of disciplines in SSH (McGillivray et al. 2020).

Despite obvious benefits, high volume and velocity social big data are not yet widely used in SSH research for at least three reasons:

1) Access to social media data requires some degree of technical competence;
2) Social media data contain a lot of noise (e.g. material from software robots), which can seriously skew results;
3) For proprietary and ethical reasons, researchers have had limited access to social background information of account holders, which restricts theoretical insight based on social media in SSH.

OUR OBJECTIVES

In collaboration with other work packages with similar interests, we want to develop a unified process for making social media data available for researchers and to develop generic tools to enrich these data.

We operate in accord with the FAIR principles (findable, accessible, interoperable and reusable) (Wilkinson et al. 2016).

Our work centers for now on Twitter, because of its rich, but underused, metadata. The future objective is to extend the work to other platforms.

This WP has four concrete objectives for 2022-23:

- To ensure that researchers have a representative and documented benchmark corpus of tweets and metadata available in real-time;
- To develop intelligent back office solutions to handle noise and enrich the user-related background information using metadata;
- To make these data freely available for fundamental research through an easy-to-use and intuitive graphic user interface that offers a user not only some basic analysis functions (e.g. a visualization tool), but also full control of the data through download functions;
- To network with other DH researchers interested in big social media data.

1. REPRESENTATIVE AND REAL-TIME DATA

We have a massive dataset of tweets from c. 500.000 user accounts from the five Nordic countries (with circa 0.5 billion words). These material have been downloaded since 2016 and the objective is to continue this "digital observatory" for several years.

The material is highly multilingual with over 70 languages represented. Before we make it freely available for fundamental research, we want to ensure that we know what it represents, how it might be skewed, fix possible problems, and document the material well.

This material can be used as primary material by researchers in various fields (cultural history, dialectology, the Nordic languages, etc.) and also as a benchmark by students and researchers collecting social media data through other tools, such as *snscrape*.

2. SMART SOLUTIONS

We develop automated back office solutions to improve data quality and to enrich social media data. First, we approach data quality from the perspective of identifying material generated by software bots, as opposed to humans. We currently have a pilot bot recognition software (Lundberg, Nordqvist & Laitinen 2019), which works reasonably well for the most obvious bot types but is unable to detect new bots having only few messages (cold-start users) and has limited capability to adapt to changing environments.

Second, the rich metadata on Twitter metadata related to interaction offer plenty of opportunities for researchers (Laitinen & Fatemi, in press).

Figure 1. A directed graph network with an ego and blue follower and black friend-follower edges

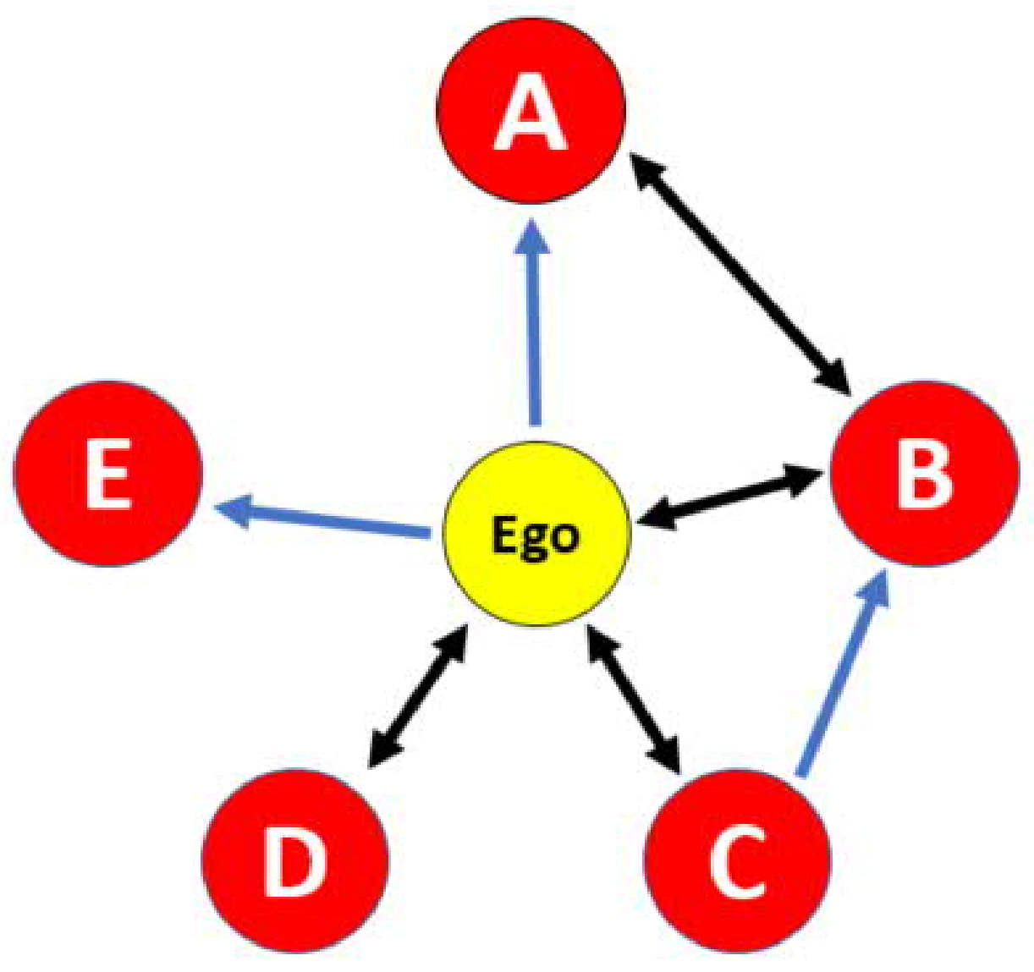


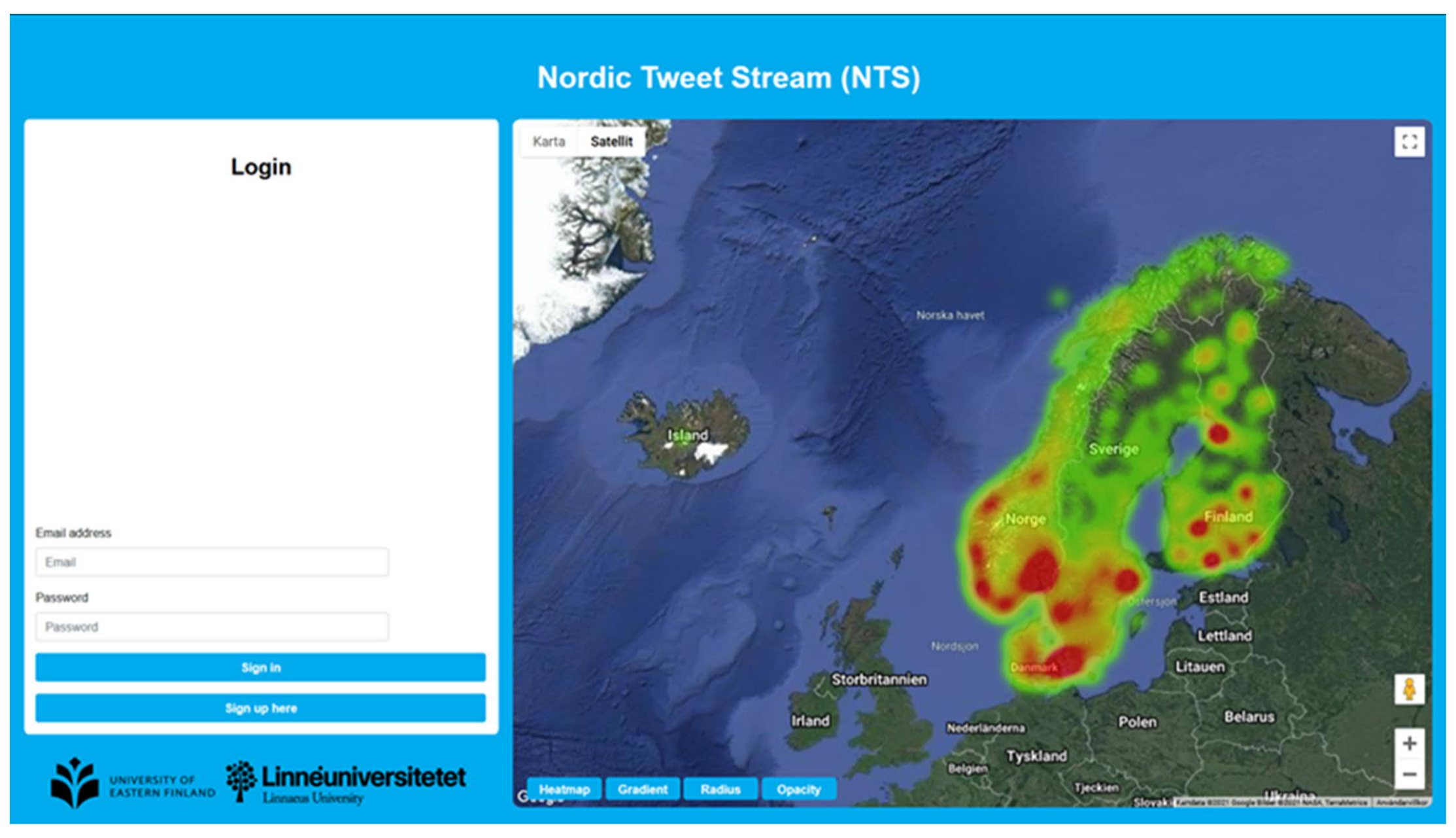
Figure 1 illustrates how interaction metadata can be used to construct a social background information based on interaction. Various previous efforts have used information provided by users themselves, making them unreliable, but interaction metadata is automatically generated and based on real interaction.

3. MAKING DATA OPEN

We plan on making this large dataset available through the Language Bank and through a user interface that has been designed for a prototypical humanities researcher. This interface is intuitive and offers some basic search functions (keyword search with basic regular expressions, geographic search, language-restricted search, any combinations thereof, etc.).

These basic functions should fulfil most needs, but to accommodate further needs, the output can also be downloaded to be processed further (ensures interoperability and reuse).

Figure 2. Login and registration page of the tweet database.


Nordic Tweet Stream (NTS)

Login

Email address
Email
Password
Password
Sign in
Sign up here

Overall, we need consistent and coordinated effort nationally and internationally to ensure that social media data remain accessible and available for fundamental research.

4. NETWORK?

We wish to partner with researchers working with other social media applications. Also, it would be userful to establish conctacts with NLP experts to train POS-taggers for noisy social media language data (Finnish and Swedish at least).

Contact: mikko.laitinen@uef.fi

REFERENCES

Laitinen, M & M. Fatemi. In press. Big and rich social networks in computational sociolinguistics.

Lundberg, J, J. Nordqvist & M Laitinen. 2019. Towards a language independent Twitter bot detector. In Navarretta et al. (eds.), *Proceedings of the 4th Digital Humanities in the Nordic Countries Conference, Copenhagen, Denmark, March 6–8, 2019*. Aachen, Germany: CEUR.

McGillivray, B. et al. 2020. The challenges and prospects of the intersection of humanities and data science: A White Paper from The Alan Turing Institute. dx.doi.org/10.6084/m9.figshare.12732164.

Wilkinson, M et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 160018.