POLICY BRIEF

Towards Better Social Intelligence for Urban Development

Intelligent Methods and Models for Mining Community Knowledge: Enabling enriched Understanding of Urban Development in Helsinki Metropolitan Region with Social Intelligence (DIGILENS-HKI) (Arcada University of Applied Sciences)

Social media is abundant with potentially valuable information to complement and enrich our understanding of our cities and society. In the meantime, it has also become clear that more and more misinformation and disinformation are generated and spread through social media. To extract meaningful social intelligence from vast amount of social media data calls for smart applications of supervised and unsupervised machine learning methods and AI techniques, to develop a rich set of tools and models for analysing the content.

Meanwhile, more rigorous evaluation and repeated validation practices are necessary to advance the field and analytical solutions. Social media data needs to be complemented by other types of relevant data to better answer decisions concerns. Key fact-checking needs to be remembered and potantial cost of errors needs to be considered when making use of social media data.

To make good use of social media information as intelligence support for urban development, planning, policy and decision making, we need to address multiple technical challenges, from data quality concerns to effective methods and tools for content analysis. The DIGILENS-HKI project investigated state of the art methods and technologies for the analysis of Instagram and Twitter data from the Hesinki metropolitan regions, and developed models and applications including topic modelling analysis, city event extraction, named entity recognition and visualization, and social media sentiment analysis.

The project also tried to raise awareness and to develop a basic understanding about bot activities and information disorder on social media. Valuable advice and practical guidance for analyzing social media data can be drawn from our research. The analytical methods and tools developed in the project are open sourced and generally applicable to social media content about any concerned topics.

Our research has benefited greatly from the recent advances in development of AI methods. The project has helped us better understand the pros and cons of the current AI methods, models and tools, the importance of testing and validation processes, as well as pre-processing choices.

In social media content analysis, a very significant part of data processing goes into preprocessing. Some pre-processing has become standarzied over time, especially with Twitter data, which can be followed and customized by practical analysis efforts. In general the nature of social media data need to be better understood and the quality of the data to be analysed need to be controlled so that the risk of drawing easy conslusions from biased social media data can be avoided.

There are large amounts of data and modelling resources available for analyzing social media content in English. However, anayitical resources for Finnish and Swedish content are much more limited, especially in sentiment analysis of Finnish text. In order to improve modelling performance on Finnish social media content, it is worthwhile to make efforts in developing labeled datasets of social media data in Finnish.

Socia media data vs Conventional data

Conventional ways for collecting data to support our understanding of cities and societies are generally considered more reliable but they are very labor intensive and often expensive. The processes can be slow, do not scale up easily and often produce data that is sparse with coarse location granularity and minimal context information.

Modern day citizens generate and share large amounts of information about where they are and what they are doing on social media, leaving marks and notes of their interaction with the urban environment and creating considerable amount of public discourse. Such social media data are sometimes biased and generally less reliable but much cheaper, easier and faster to collect in massive amounts as timely geo-tagged data with fine-grained location data, rich demographics and more context information.

Social media could host potentially very valuable information and community knowledge. In the meantime, it has also become clear that more and more misinformation and fake content are intentionally generated and spread through social media. The massive spread of digital misinformation has been identified as one of the major global risks. To understand the nature of information disorder on social media, we conducted literature and media study to present an overview of bot activities and information disorder to guide the practical use of social media data.

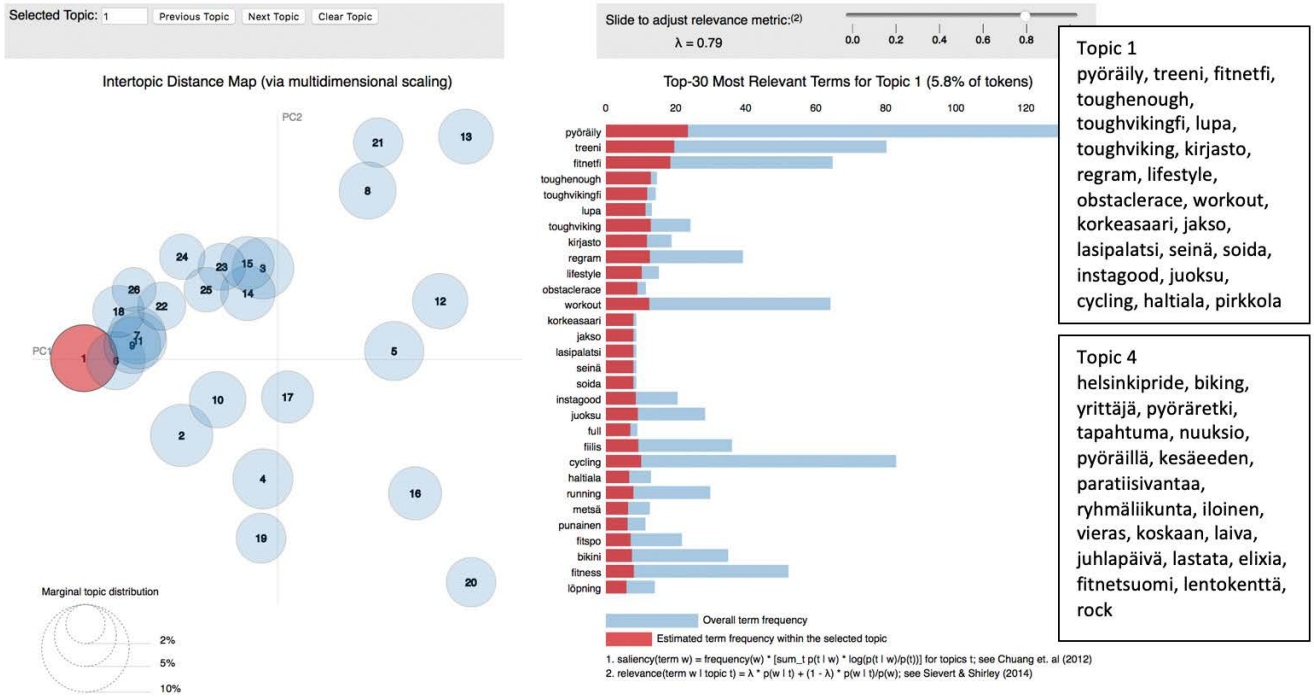
Topic modelling analysis, named entities recognition and sentiment analysis

Working on Instagram and Twitter data from the Helsinki region, the project explored state of the art natural language processing techniques, as well as machine learning and AI methods for analysing social media content data. Our study helps to open up a good understanding of the possibilities and means for the using of social media data in, for example, urban planning and urban development.

To understand the topic content in large collection of social media posts and discussions, we developed topic modelling analysis and visualization tools to help explore the presence and prevalence of selected topics in social media, such as festival events, cycling and transportation, safety issues, and so on. In collaboration with the Digital Georgraph Lab of Helsinki University, we explored the application of our models and tools to analyzing cycling related topics from the Instagram data for summer 2016 in English and Finnish languages (Figures).

We addressed the challenging task of emerging Named Entity Recognition from user generated noisy content. Applying recurrent neural networks and topic modelling methods, we developed deep neural models and visualization tool for analyzing named entities. We also developed deep neural models for social media sentiment analysis, applying deep learning and transfer learning methods.

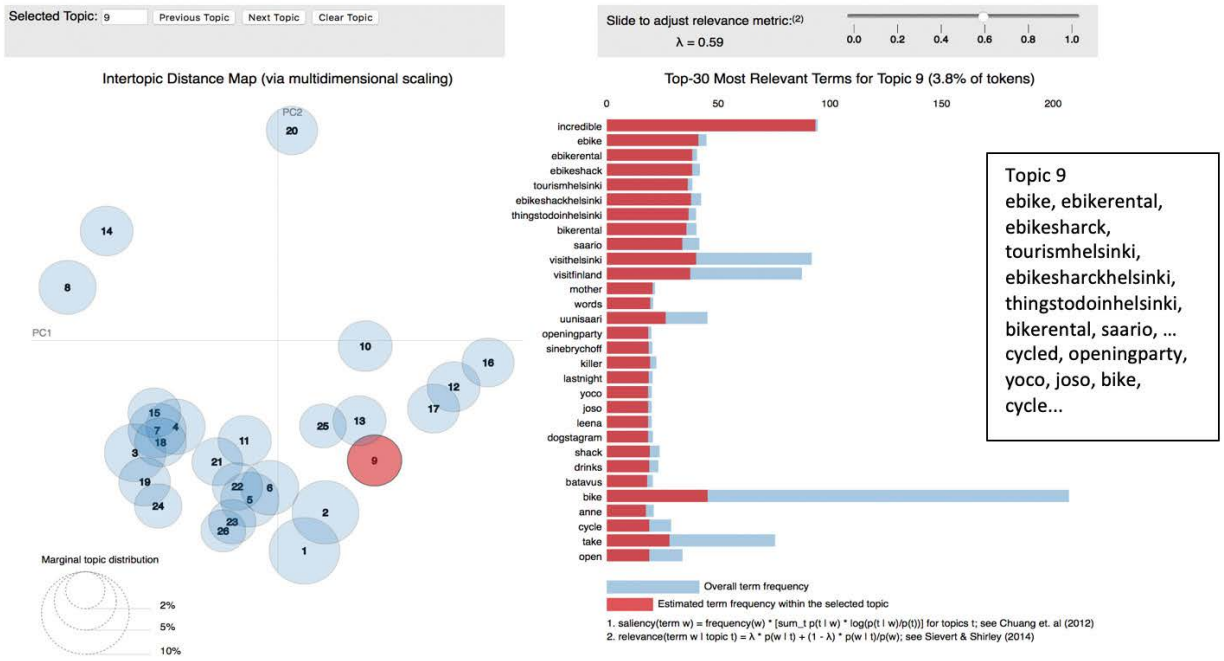
Cycling related topics (Posts in Finnish), 26 topics LDA model



Topic 1
pyöräily, treeni, fitnetfi, toughenough, toughvikingfi, lupa, toughviking, kirjasto, regram, lifestyle, obstaclerace, workout, korkeasaari, jakso, lasipalatsi, seinä, soida, instagood, juoksu, cycling, haltiala, pirkkola

Topic 4
helsinkipride, biking, yrittäjä, pyöräretki, tapahtuma, nuuksio, pyöräillä, kesäeeden, paratiisivantaa, ryhmäliikunta, iloinen, vieras, koskaan, laiva, juhlapäivä, lastata, elixia, fitnetsuomi, lentokenttä, rock

Cycling related topics (Posts in English), 26 topics LDA model



Advice on the use of AI Tools in Social Media Data Analysis

Topic Modelling Analysis with data visualization offers a simple yet powerful means for exploring large amounts of social media content. It helps to discover specific topics and events with natural granularity. Its unsupervised nature makes it easy to apply in practice.

Named Entity Recognition and high-dimensional dataset visualization can provide a quick overview of names of people, places, organizations, products and creative works, for example. However, emerging named entity recognition analysis of social media data is still very challenging and includes many problems far from being solved. The performance levels with the available tools are still unsatisfactory in general.

Large amounts of analytical resources for social media data analysis work best for content in English. This would mean performance drop from state of the art results when analyzing Finnish and Swedish content. Post-processing can often help bring improvements of analytical results. The post-processing rules are formulated and accumulated from testing and validation processes.

Social media data preprocessing choices can also have positive or negative effects on the analysis results. Fortunately with much efforts from the research community, the preprocessing components are becoming standardized and good to be used as reference when developing applications.

The importance of validation can not be over emphasized. In developing practical analytical applications, involvement of domain experts such as, in this case, planners or decision-makers, is imperative for quality control of the analytical process.

Before considering using social media content data as information for decision-making, it is extremely important to understand the nature and quality of the social media platforms and datasets in each case. It is fundamental to understand the effects of biases and information disorder on social media content as well as biases in the analytical methods. It is also important to maintain the awareness of the traning datasets limitation in size and time span. Larger complmentary data sources needs to be incorporated to support real world decision making.

Proposal for Action

One critical restriction for developing better performing analytical tools is the limited availability of suitable training datasets. In order to improve analytical results for social media content in Finnish language, it is important to make more efforts in developing more and larger labeled datasets. This could be a joint-effort of the research and user community and the process could be crowdsourced. The sooner we have such larger labelled datasets available, the sooner the analysis of social media content in Finnish will improve.

Meanwhile, research work is emerging on better learning methods for low resource languages. Progress in this area can hopefully bring us more insights and resources for processing and analyzing Finnish data as well. This is an important direction for following up research and development of applications.

When using social media data as inputs to important decision-making and policy processes, remember key fact-checking by using information external to the social media data. The analytical results need always to be validated and conclusions drawn with caution.

Publications:

Shuhua Liu and Patrick Jansson, "Topic Modelling Analysis of Instagram Data for the Greater Helsinki Region", Arcada Working Paper 3/2017, Arcada University of Applied Sciences https://www.theseus.fi/bitstream/handle/10024/140608/AWP_3_2017_Topic_modelling.pdf?sequence=1&isAllowed=y .

Patrick Jansson and Shuhua Liu, "Topic Modelling enriched LSTM Models for the Detection of Novel and Emerging Named Entities from Social Media", SocialNLP 2017 the 5th International Workshop on Natural Language Processing for Social Media, December 11, 2017, Boston. https://ieeexplore.ieee.org/document/8258462

Shuhua Liu and Patrick Jansson, "City Event Detection from Social Media with Neural Embeddings and Topic Model Visualisation", IWSC 2017 the 3rd International Workshop on Smart Cities: People, Technology, and Data, December 11, 2017, Boston. https://ieeexplore.ieee.org/document/8258430

Shuhua Liu, "Bot Activities and Information Disorder on Social Media", Arcada Working Paper 2019, Arcada UAS

More information:

Dr. Shuhua Liu, Senior Research Fellow
shuhua.liu@arcada.fi

Project website:
https://rdi.arcada.fi/katumetro-digilens-hki

Github repository:
https://github.com/ShuhuaLiu/DIGILENS-HKI

Partner:
Digital Geography Lab, Helsinki University

Helsinki Metropolitan Region Urban Research Program

(2010-2018) is a horizontal cooperation network between Helsinki metropolitan area cities, universities, universities of applied sciences and two state ministries. Main goal of the program is to promote and fund multi-disciplinary, high quality urban research with a starting point that takes into consideration the special characteristics of the Helsinki Metropolitan Region.

The program aims to provide up to date scientific research results, data and practical knowledge as the basis of decision-making, to create best-practices and to help generate new innovative operational models of cooperation between different actors in the region. Special attention is paid on dissemination and improving usability of research data. The program funds demand-based urban research projects and development activities on current topics and issues in the Helsinki metropolitan area.

metropolitutkimus.fi

Project Partners:

ARCADA

KAUPUNKITUTKIMUS JA METROPOLIPOLITIIKKA