Sentiments in Helsinki - Spatiotemporal Analysis of Instagram Posts

Qazi Firas | Tuomo Hiippala | Iuliia Kim | Anton Matveev | Sid Rao | Saara Suominen | Tuuli Toivonen | Elias Willberg
What is sentiment?
Research questions

1. **Spatial** - How sentiment polarity is distributed in the neighborhoods of Helsinki?

2. **Temporal** - What is the variation of sentiments over time?
Data - What did we have?

- **1,316,705** Instagram posts.
- **Time**: 1\textsuperscript{st} of June 2014 to 31\textsuperscript{st} of March 2016
- **Location**: Helsinki Metropolitan Area

Posts within Helsinki, that are in English: **193,111**
Process Outline - Our plan

**Top Priority**
- Data cleaning
- Language identification
- Sentiment analysis
- Use GIS to make maps

**Back-Burner**
- Topic modeling
- Named Entity Recognition
- Computer Vision analysis
Step 1: Preprocessing

Cleaning the data by:

- Removing posts with no caption.
- Removing posts with no text (containing only emojis and hashtags).

Filter by restricting the posts to only those are:

- Within Helsinki;
- In English language.
Step 2: Language detection

- Available options:
  - Langdetect (55 languages)
  - Langid (97 languages)
  - Also, NLTK
  - FastText

- We chose: FastText
  - Pre-trained language identification models for 176 languages.
  - Very fast and reliable
  - State-of-the-art library by Facebook Research
    - Suitable for Instagram and other social media.
Step 3: Sentiment analysis

- **Used tools:**
  - VADER (analyze clear text without hashtags and emojis)
  - Aylien API (analyze whole captions)
  - Checked against **manually annotated gold standard**.

- **Filtering results:**
  - set threshold of polarity confidence to 0.7

- **Obstacles:**
  - hashtags are inserted into sentences and should be considered as their integrated part
Sentiment analysis

3 - positive
2 - neutral
1 - negative

text + hashtags + emojis
- text + emojis
- text + hashtags
- clear text
Plotting the data on the map

Dividing Helsinki into discernible units.

Considered options:

- Postcode division
- **Neighborhoods**
- Square grids
- Land use
Some of the results:

- Raw Instagram data is tough to process
- A noticeable positive-sentiment skew
- User activity peaks during winter and goes down in summer
- The city center is generally more positive
Limitations & problems

Common problems of working with geotagged SoMe data:

- Accessibility: API no longer working -> data is not recent
- Language usage: slang, codeswitching
- Pictures not accessible

Other:

- Named Entity Recognition was not accurate.
- Language detection may be not so accurate.
Limitations: Negative sentiment on social media
Ideas for future research

1. To employ topic modeling to the posts in different neighborhoods.

2. To compare the results to other kinds of geographical data: land use maps, levels of income etc.

3. To extract only the strongly positive posts, and study the topics that occur in them.

4. To study the pictures as well.

5. Close reading and case studies in addition to quantitative methods.
Thank You