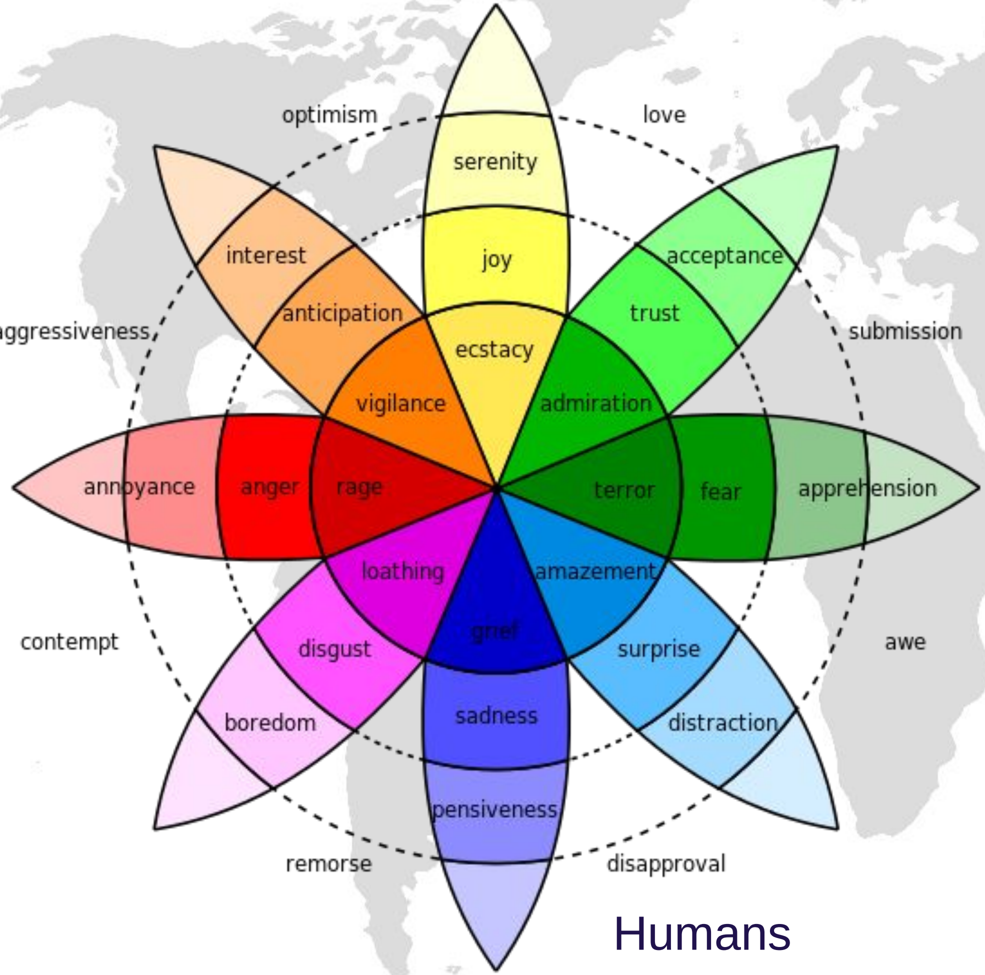
Sentiments in Helsinki - Spatiotemporal Analysis of Instagram Posts

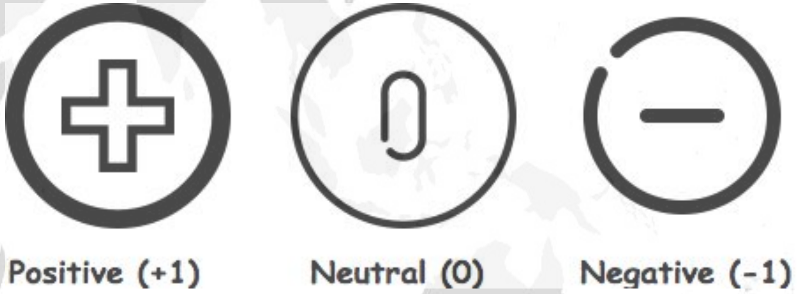
Qazi Firas | Tuomo Hiippala | Iuliia Kim | Anton Matveev | Sid Rao | Saara Suominen | Tuuli Toivonen | Elias Willberg

What is sentiment?



Humans

Computers

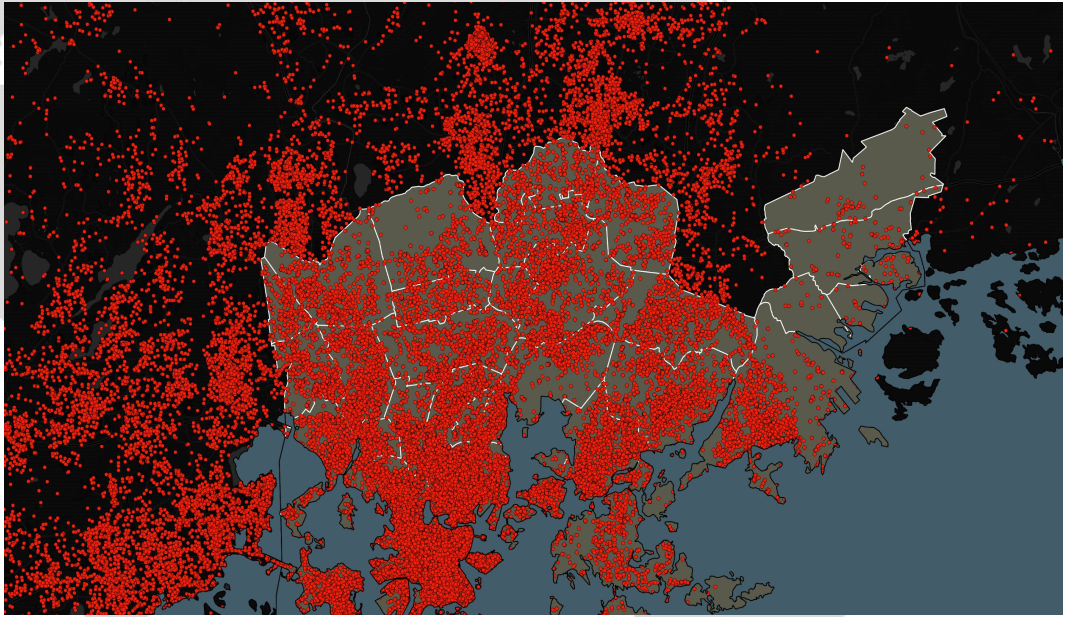


Research questions

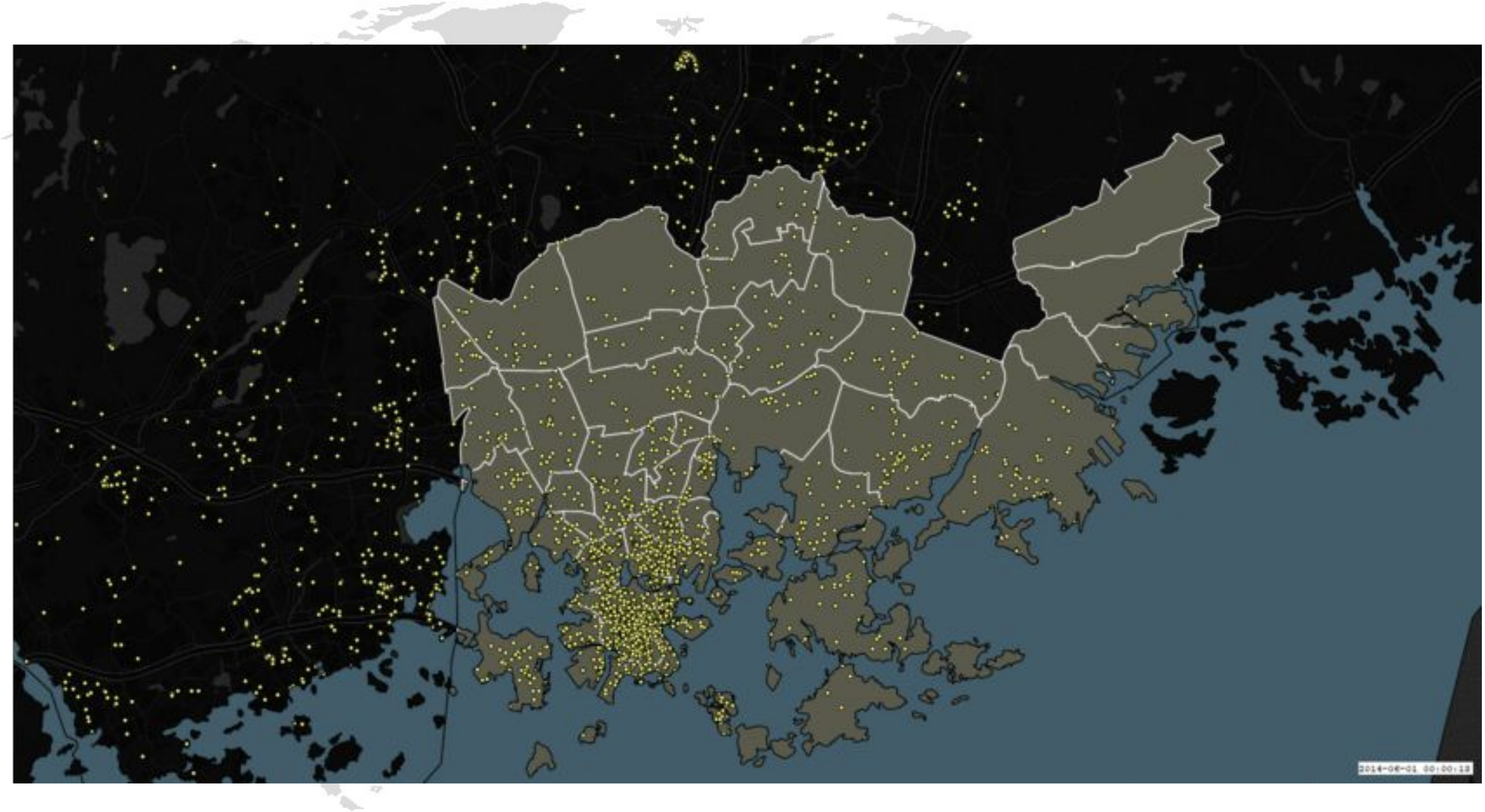
1. Spatial - How sentiment polarity is distributed in the neighborhoods of Helsinki?
2. Temporal - What is the variation of sentiments over time?

Data - What did we have?

- **1,316,705** Instagram posts.
- **Time**: 1st of June 2014 to 31st 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.

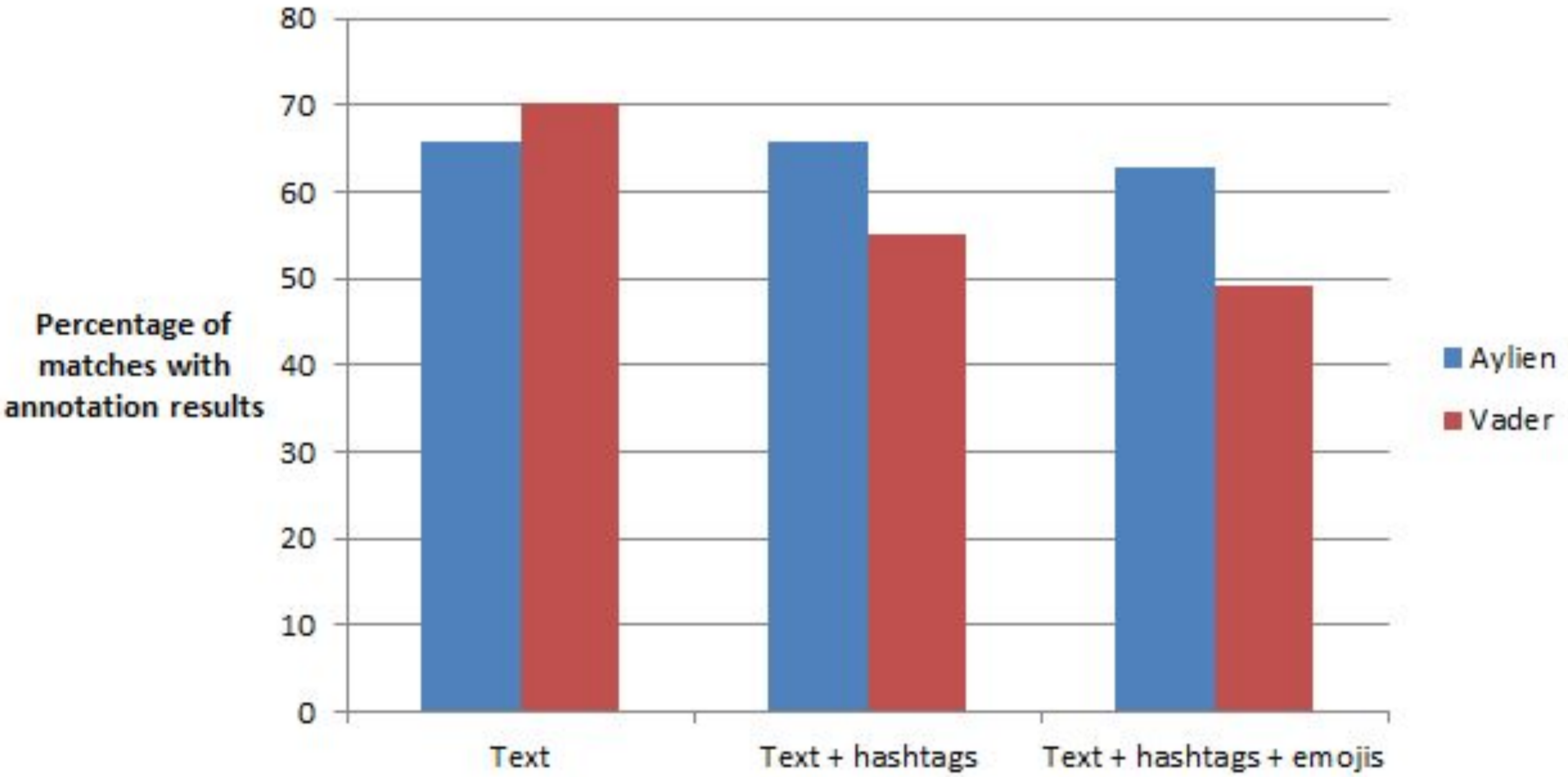
Step 2: Language detection

- 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.


IT+ FI, 508
UK, 549
TR, 840
TH, 1084
EN+RU, 1086
ET, 1735
FR, 1950
AR, 2140
PT, 2234
IT, 2378
ZH, 2399
DE, 2869
ES, 3068
KO, 3621
JA, 8145
EN+ FI, 11864
SV, 13344
RU, 45118
EN, 293063
FI, 317901

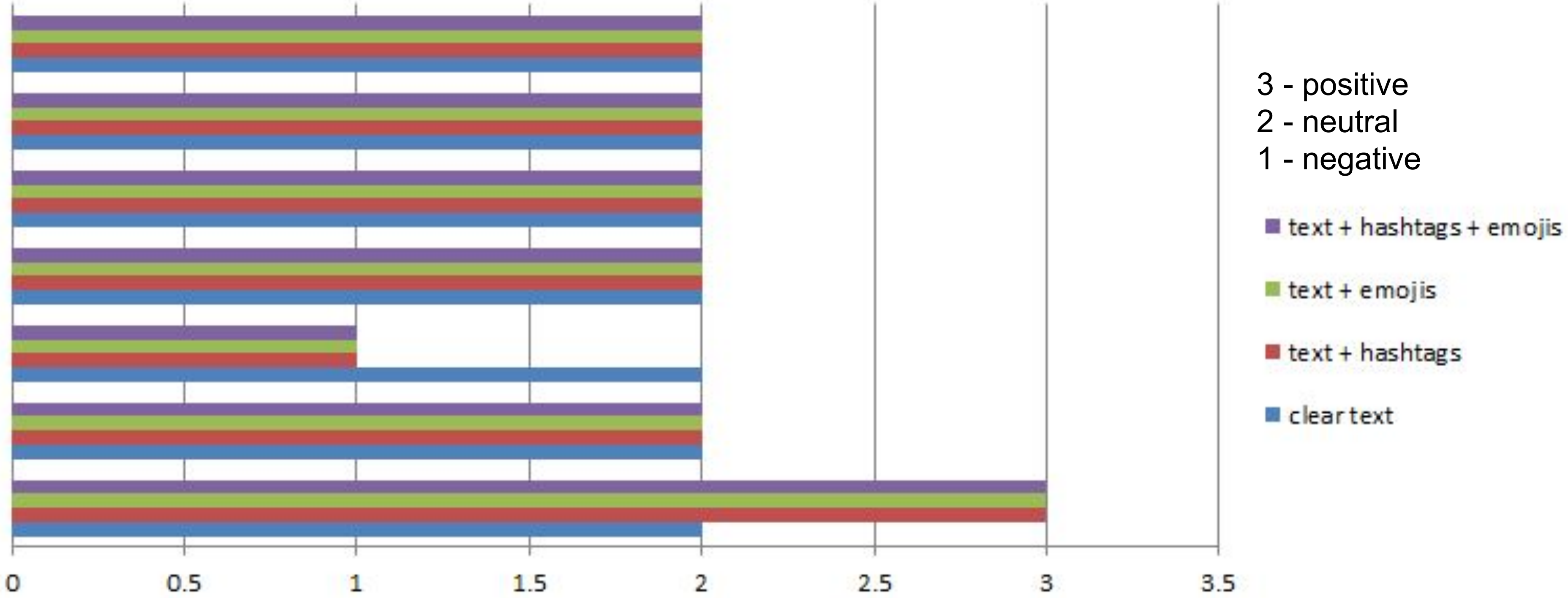
Step 3: Sentiment analysis

- 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

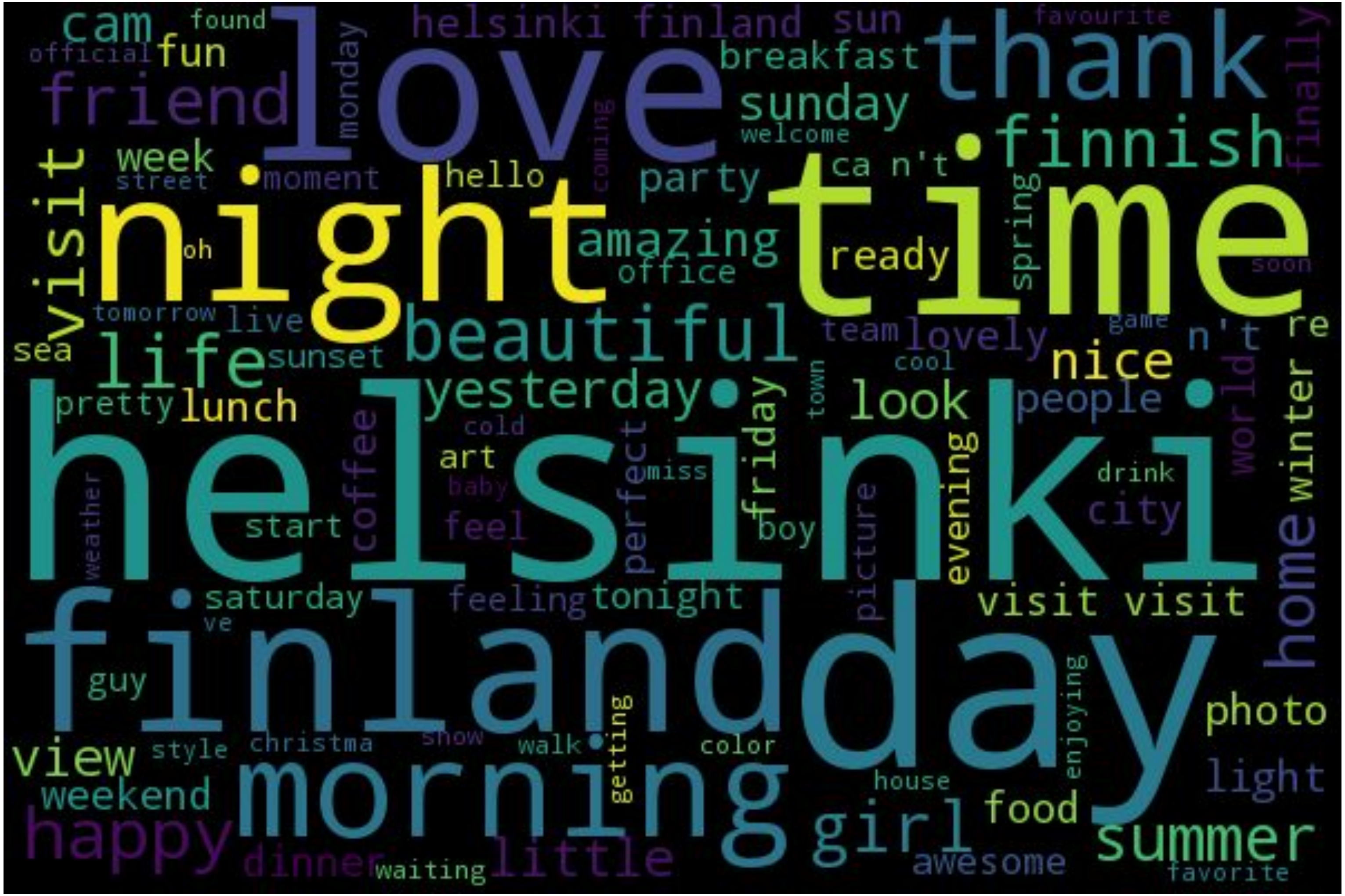


Percentage of matches with annotation results
80
70
60
50
40
30
20
10
0
Text
Text + hashtags
Text + hashtags + emojis
Aylien
Vader

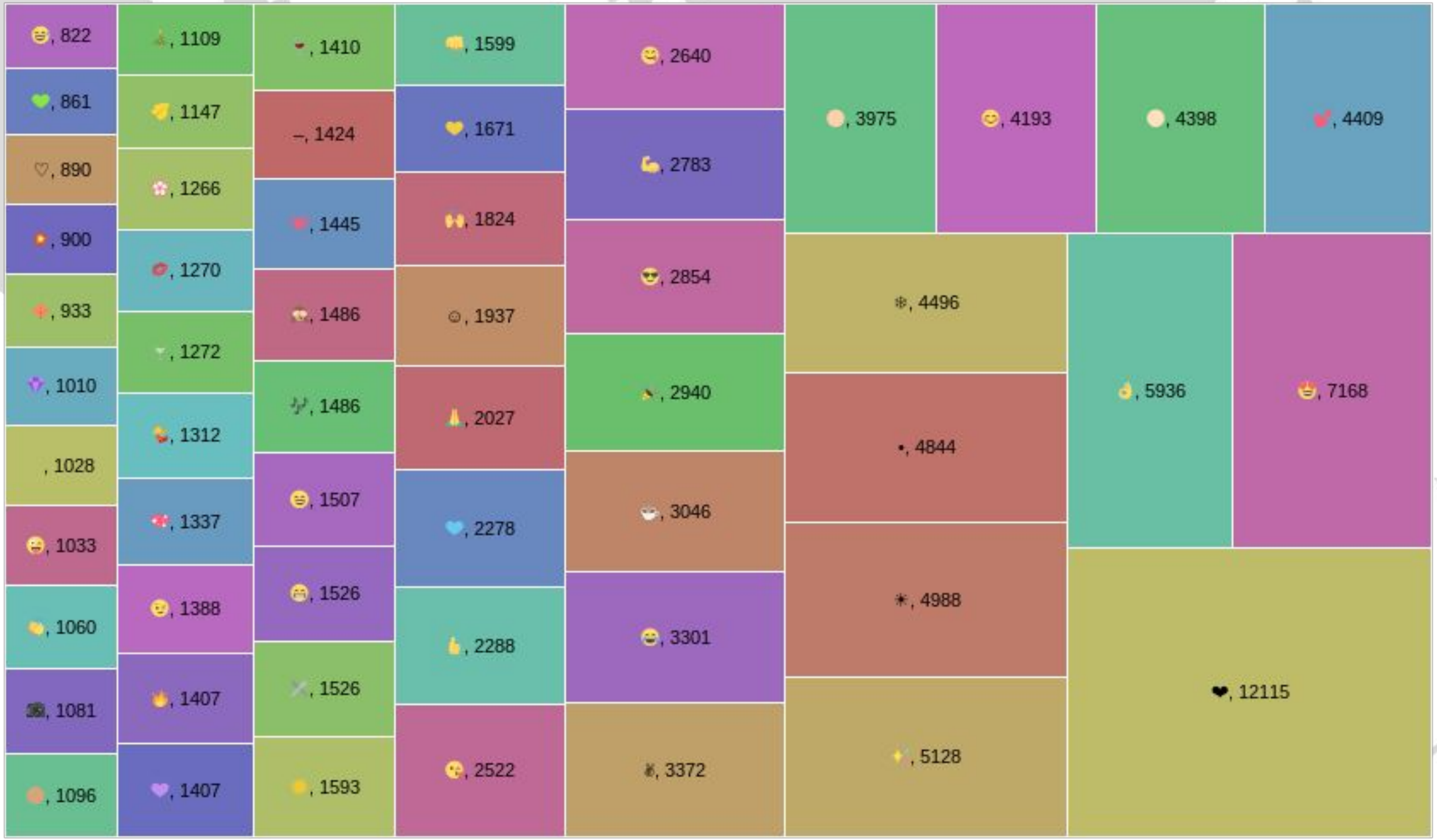
Sentiment analysis



3 - positive
2 - neutral
1 - negative


cam
found
love
helsinki finland
sun
thank
favourite
official
fun
breakfast
friend
monday
sunday
coming
welcome
finnish
finally
visit
week
street
moment
hello
party
ca n't
night
time
amazing
ready
spring
oh
office
soon
tomorrow
live
beautiful
team
lovely
game
n't
re
sea
life
sunset
cool
nice
yesterday
look
people
pretty
lunch
town
world
winter
helsinki
cold
coffee
art
perfect
friday
miss
evening
drink
baby
city
weather
start
feel
boy
picture
home
saturday
feeling
tonight
visit visit
ve
finland
day
guy
photo
christma
show
walk
getting
color
enjoying
style
view
house
light
weekend
morning
food
happy
girl
summer
dinner
waiting
little
awesome
favorite

Emoji usage



, 822
💚, 861
♡, 890
, 900
, 933
, 1010
, 1028
, 1033
, 1060
, 1081
, 1096
, 1109
, 1147
🌸, 1266
💋, 1270
, 1272
, 1312
, 1337
, 1388
🔥, 1407
💜, 1407
, 1410
–, 1424
, 1445
🙈, 1486
, 1486
, 1507
, 1526
, 1526
, 1593
👊, 1599
💛, 1671
🙌, 1824
☺, 1937
🙏, 2027
💙, 2278
👍, 2288
😘, 2522
😋, 2640
💪, 2783
😎, 2854
, 2940
, 3046
😂, 3301
, 3372
, 3975
😊, 4193
, 4398
💕, 4409
❄, 4496
•, 4844
☀, 4988
✨, 5128
👌, 5936
🤩, 7168
❤, 12115

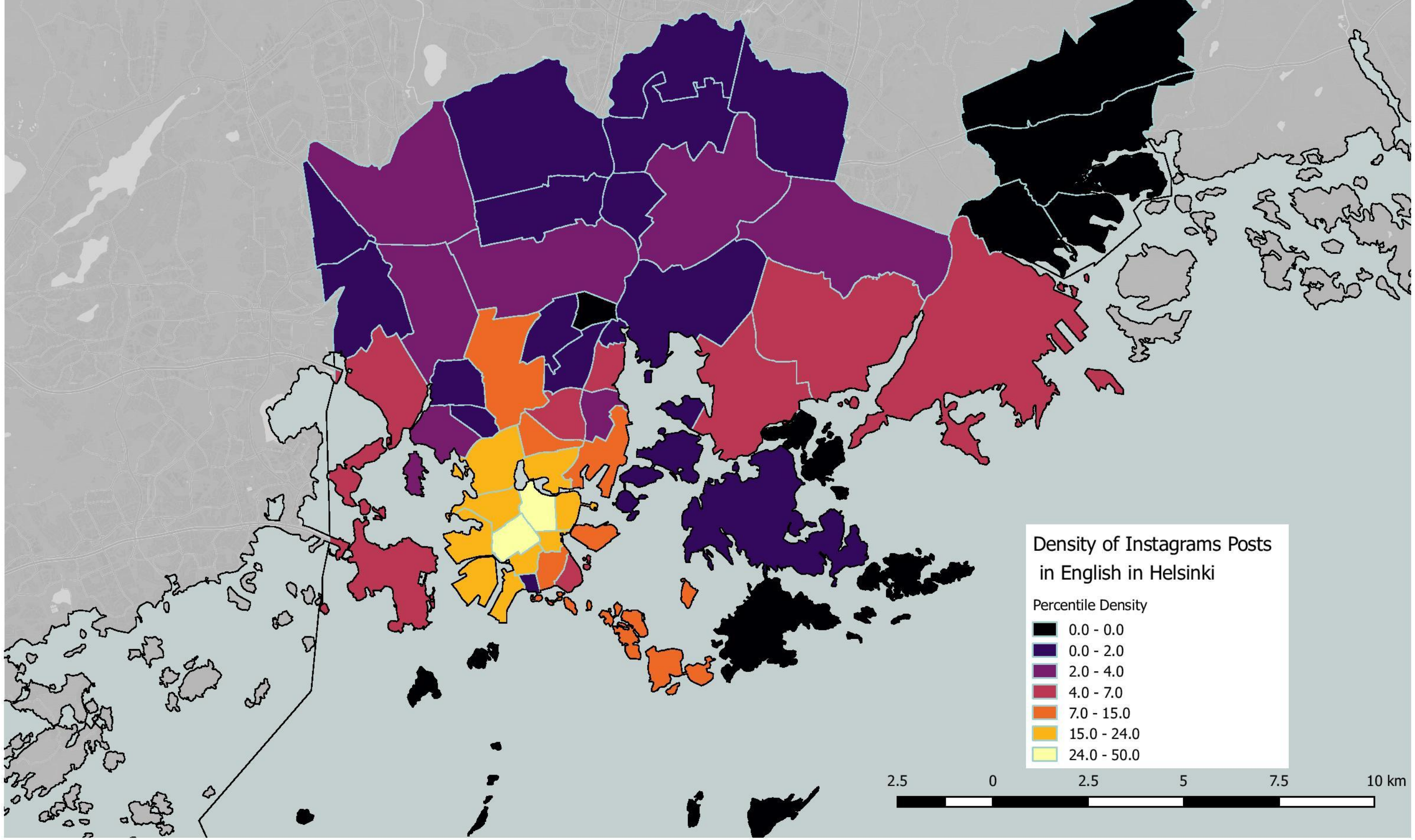
Plotting the data on the map

Dividing Helsinki into discernible units.

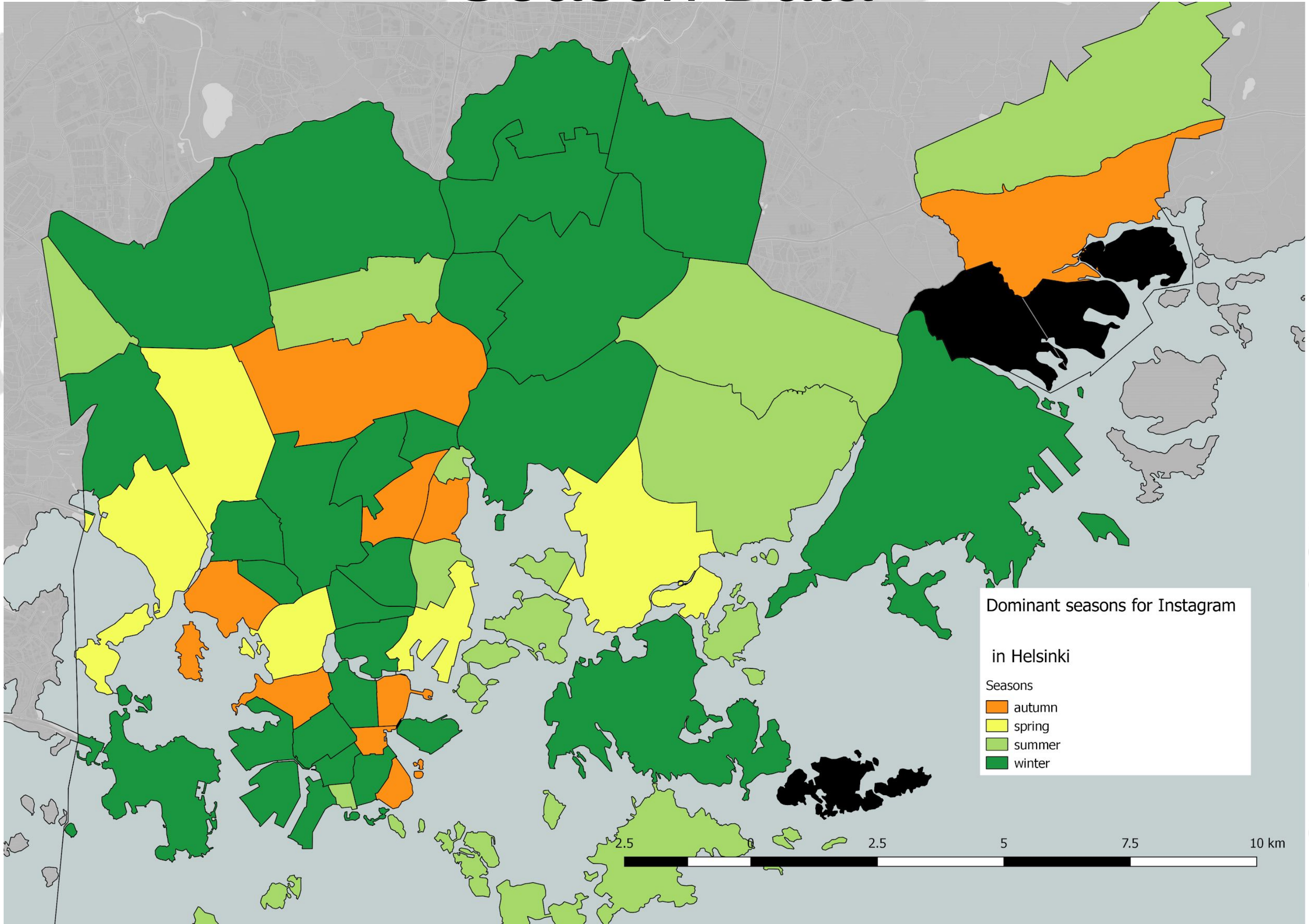
Considered options:

- Postcode division
- **Neighborhoods**
- Square grids
- Land use

Density of Posts

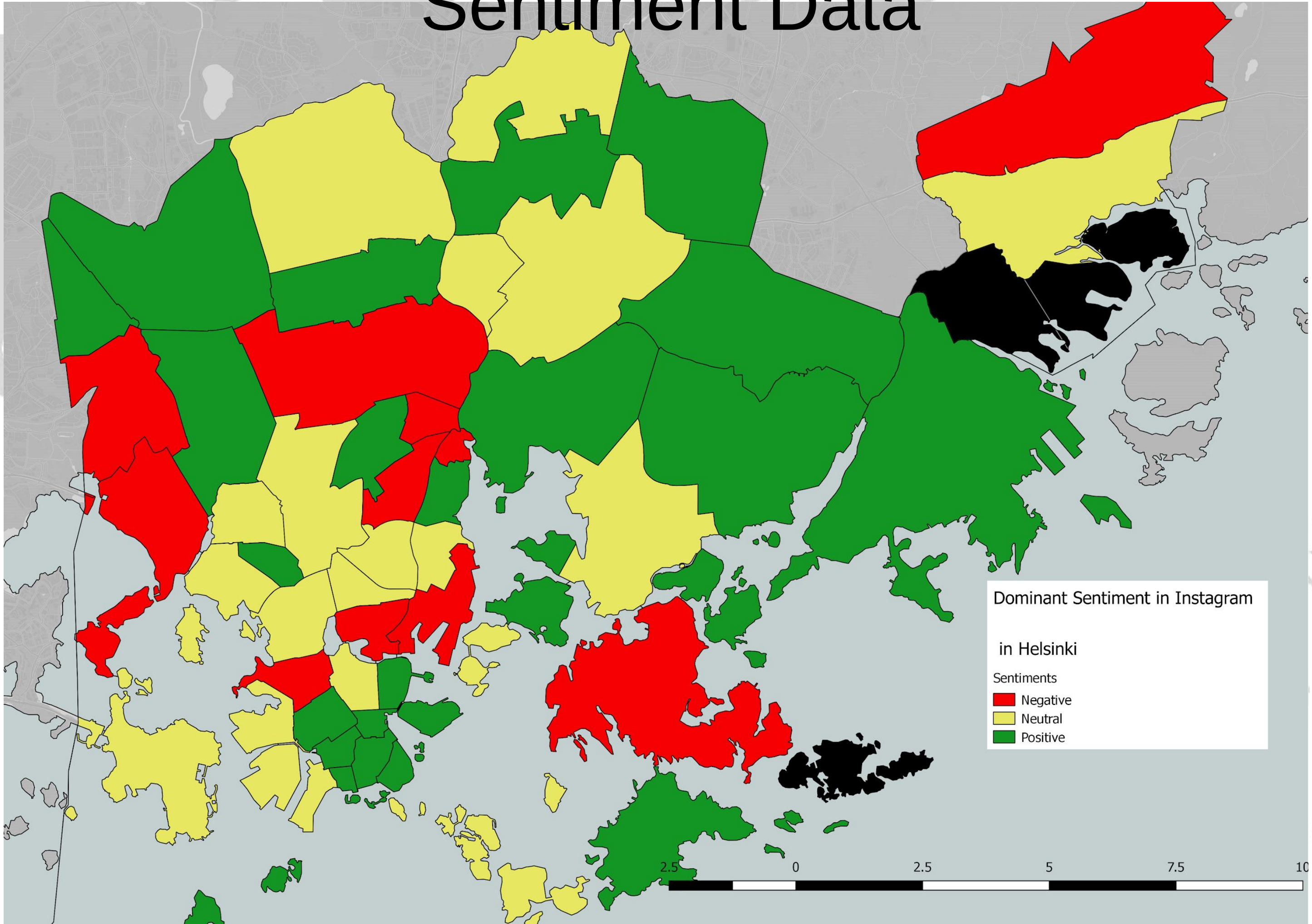


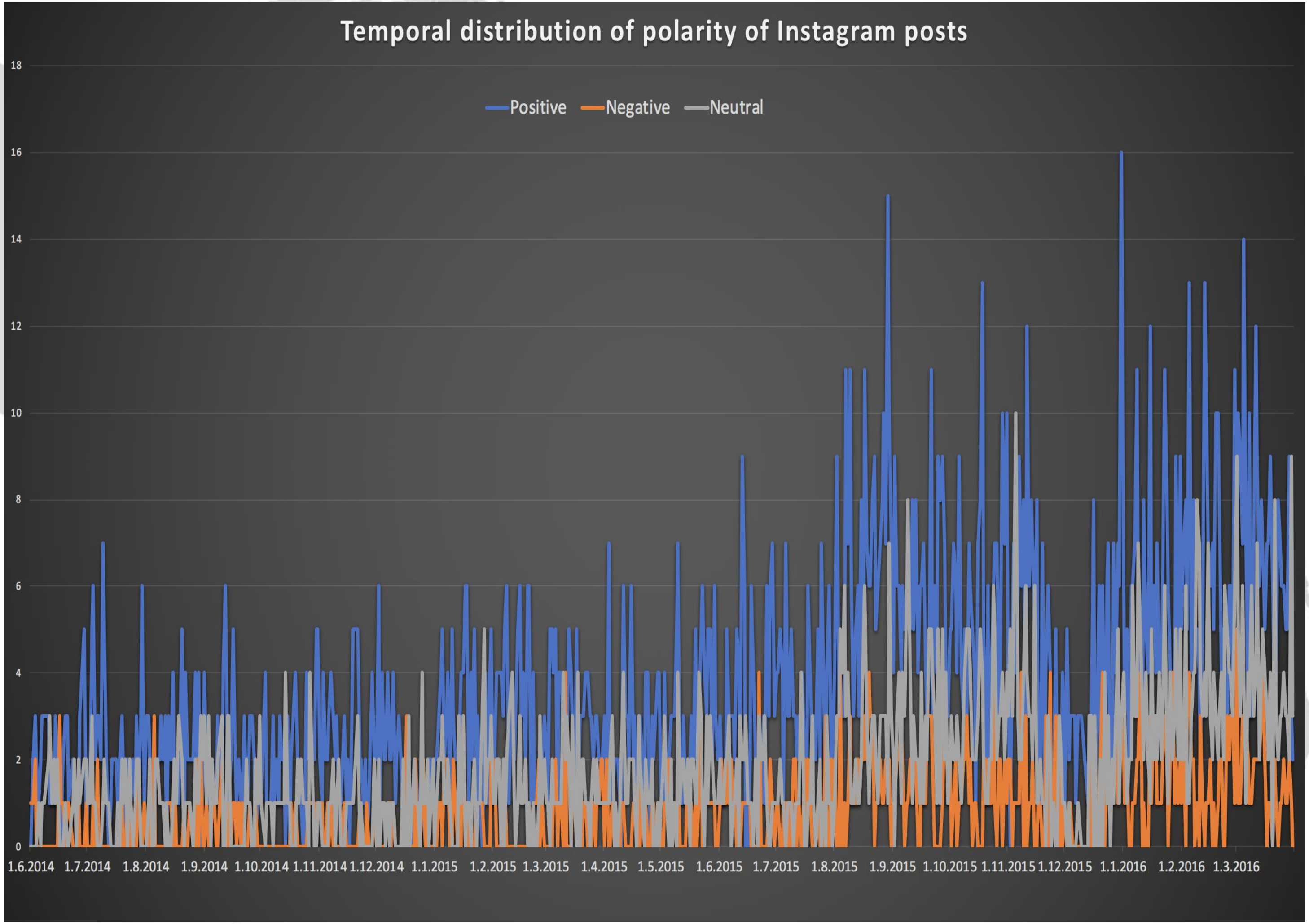
Season Data



Dominant seasons for Instagram
in Helsinki
Seasons
autumn
spring
summer
winter
2.5
0
2.5
5
7.5
10 km

Sentiment Data





Temporal distribution of polarity of Instagram posts
Positive
Negative
Neutral
18
16
14
12
10
8
6
4
2
0
1.6.2014
1.7.2014
1.8.2014
1.9.2014
1.10.2014
1.11.2014
1.12.2014
1.1.2015
1.2.2015
1.3.2015
1.4.2015
1.5.2015
1.6.2015
1.7.2015
1.8.2015
1.9.2015
1.10.2015
1.11.2015
1.12.2015
1.1.2016
1.2.2016
1.3.2016

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

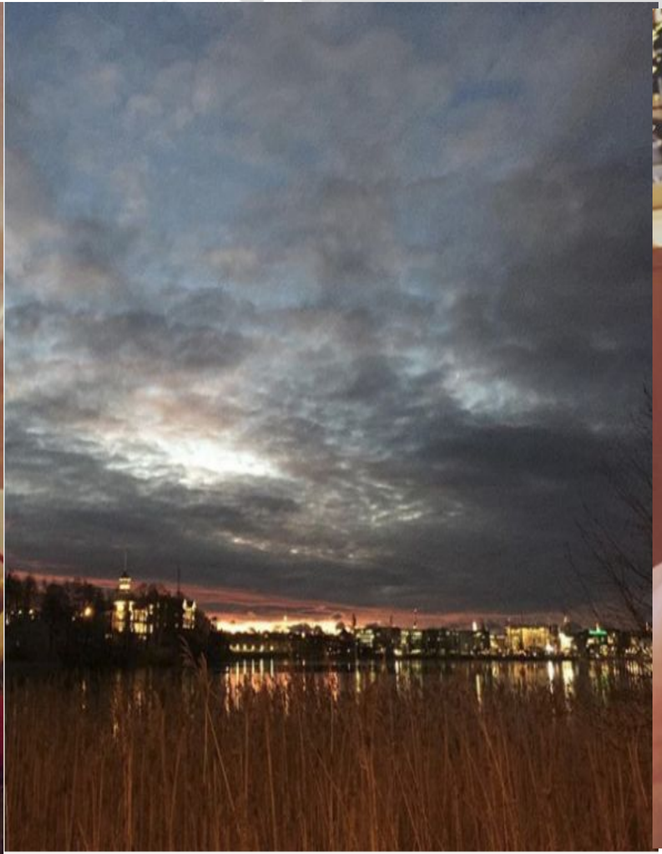
Common problems of working with geotagged SoMe data:

- 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