Cartography M.Sc.

Sentiment-based spatial-temporal event detection in social media data

Lin Che

Supervisors: Juliane Cron, Ruoxin Zhu
Reviewer: Dr. Eva Hauthal

Outline

- Introduction
- Sentiment analysis
- Spatial-temporal analysis-based event detection
- Spatial-temporal analysis-based event interpretation
- Conclusion and outlook

Introduction: Social media

Introduction: Hypothesis

The hypothesis guiding this research is that social events change population sentiment orientation (PSO) in the dimension of time and space.

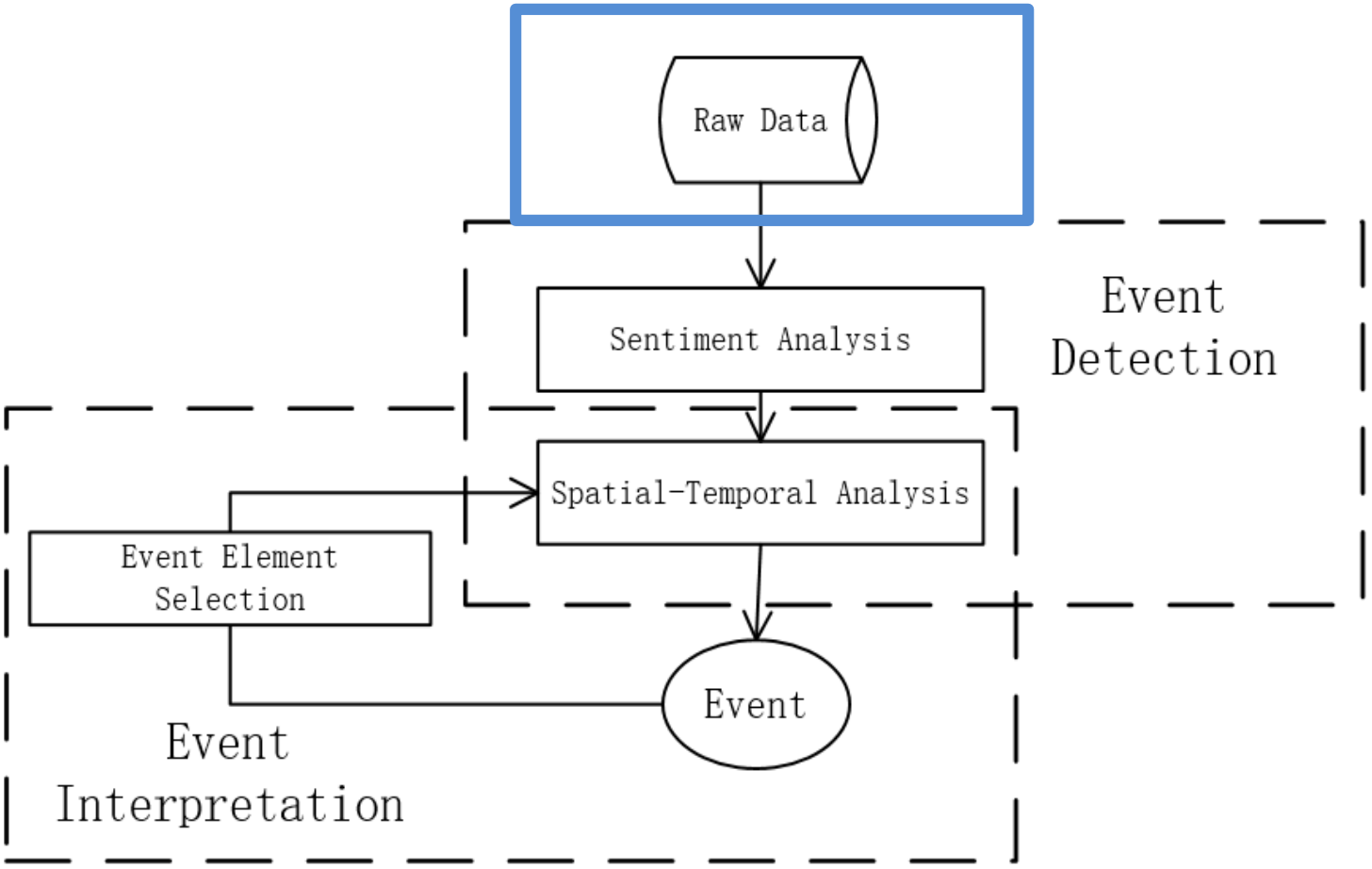
Introduction: Research objectives

- Find a suitable sentiment indicator to represent the population sentiment orientation.
- Develop a methodology for spatial-temporal analysis of the population sentiment to detect sentiment fluctuation and abnormalities.
- Find suitable methods to identify and interpret the event in the spatial-temporal dimensions.

Introduction: Research questions

- Which sentiment analysis method for the given social media is most suitable one for the PSO indicator?
- How should the spatial-temporal analysis be designed for population sentiments? Which specific methods or algorithms can be applied?
- How can events be extracted and identified based on spatial-temporal analysis? What method(s) is (are) suitable to interpret the event?

Workflow


Raw Data
Sentiment Analysis
Spatial-Temporal Analysis
Event Element Selection
Event
Event Detection
Event Interpretation

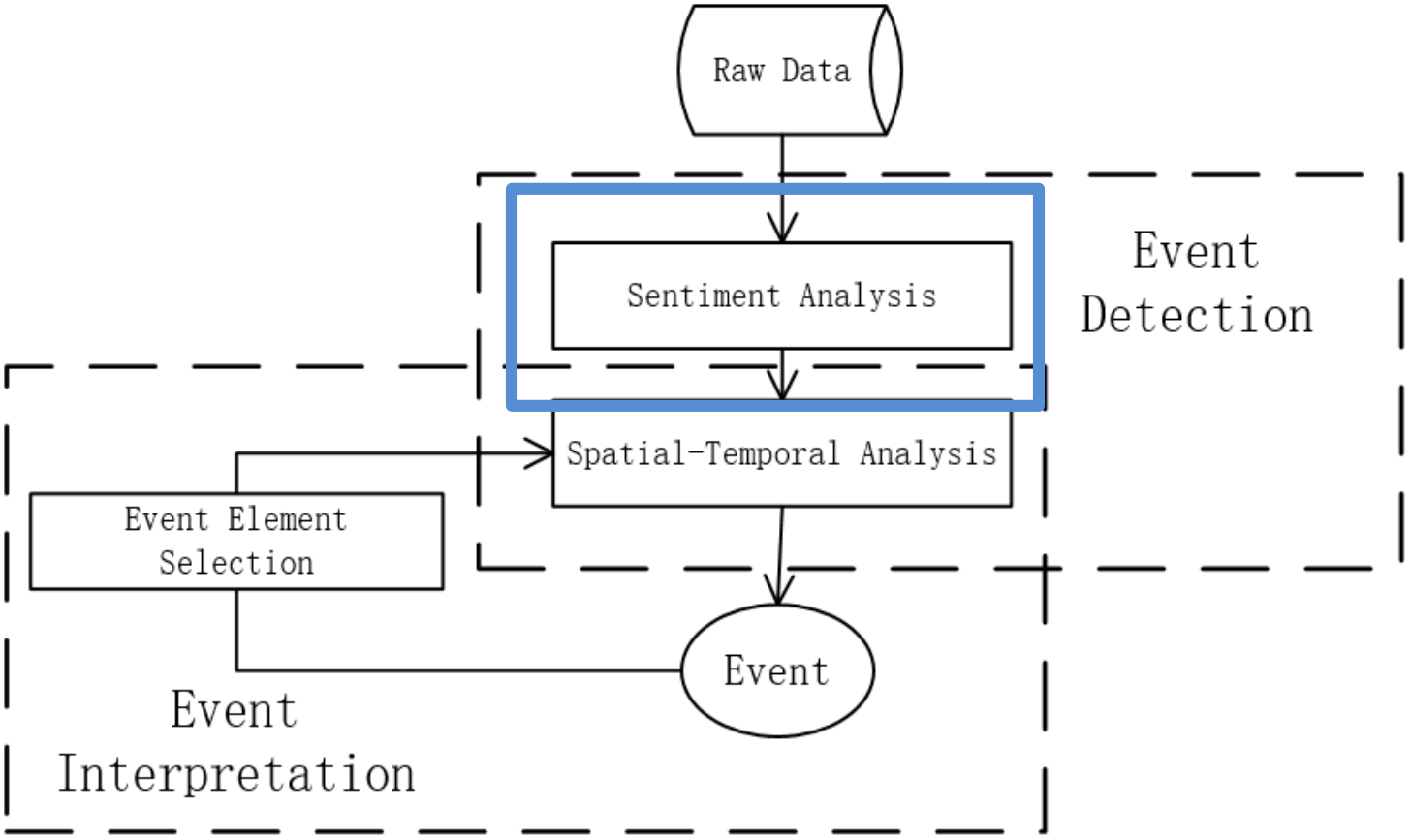
Data description

Case study data set

Sina Weibo

Attributes	Values
Source	Michael Jendryke (WHU)
Time	01.2014 – 05.2015
Region	Shanghai, China
Size	4GB
Format	CSV
Columns	28
Rows	11794009

Workflow


Raw Data
Sentiment Analysis
Event Detection
Spatial-Temporal Analysis
Event Element Selection
Event
Event Interpretation

1. Sentiment analysis

Sentiment analysis: Population sentiment orientation indicator

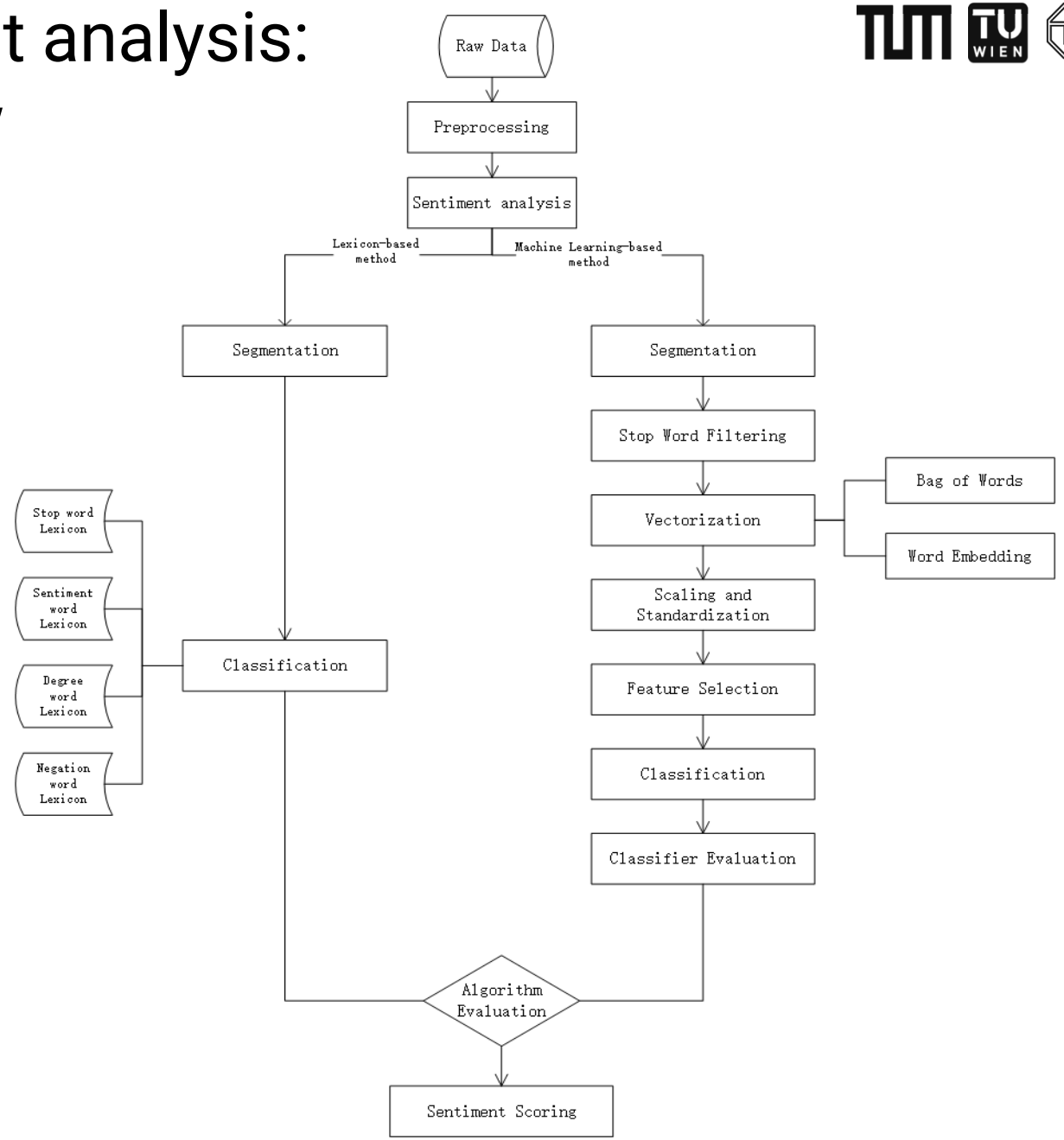
$$\mathrm{PSO} = \frac{P}{N}$$

P = number of positive records
N = number of negative records

Ground truth data set:

100k (50k positive, 50k negative)

Sentiment analysis: Workflow



Sentiment analysis: Data preprocessing

Data selection

- Data attributes selection
- Data range selection

Data attributes after selection

Attribute	Type	Description	
Created_at		String	UTC time when this Tweet was created.
id		Int64	The integer representation of the unique identifier for this Tweet.
user		User Object	The user who posted this Tweet.
text		String	The actual UTF-8 text of the status update.
Coordinates		Coordinates	Nullable. Represents the geographic location of this Tweet as reported by the user or client application. The inner coordinates array is formatted as geoJSON (longitude first, then latitude).

Sentiment analysis: Data preprocessing

Data cleaning

- Integrity check
- Deduplication check

The data types need to be removed

Data type	Example
Link	http://abcde/efg
Emoji	😊 (U+1F60A)
Punctuation	! , . ? ~
White space	" "
Digits	12345
Tag	#News#

Sentiment analysis: Lexicon-based method flow

- Text segmentation

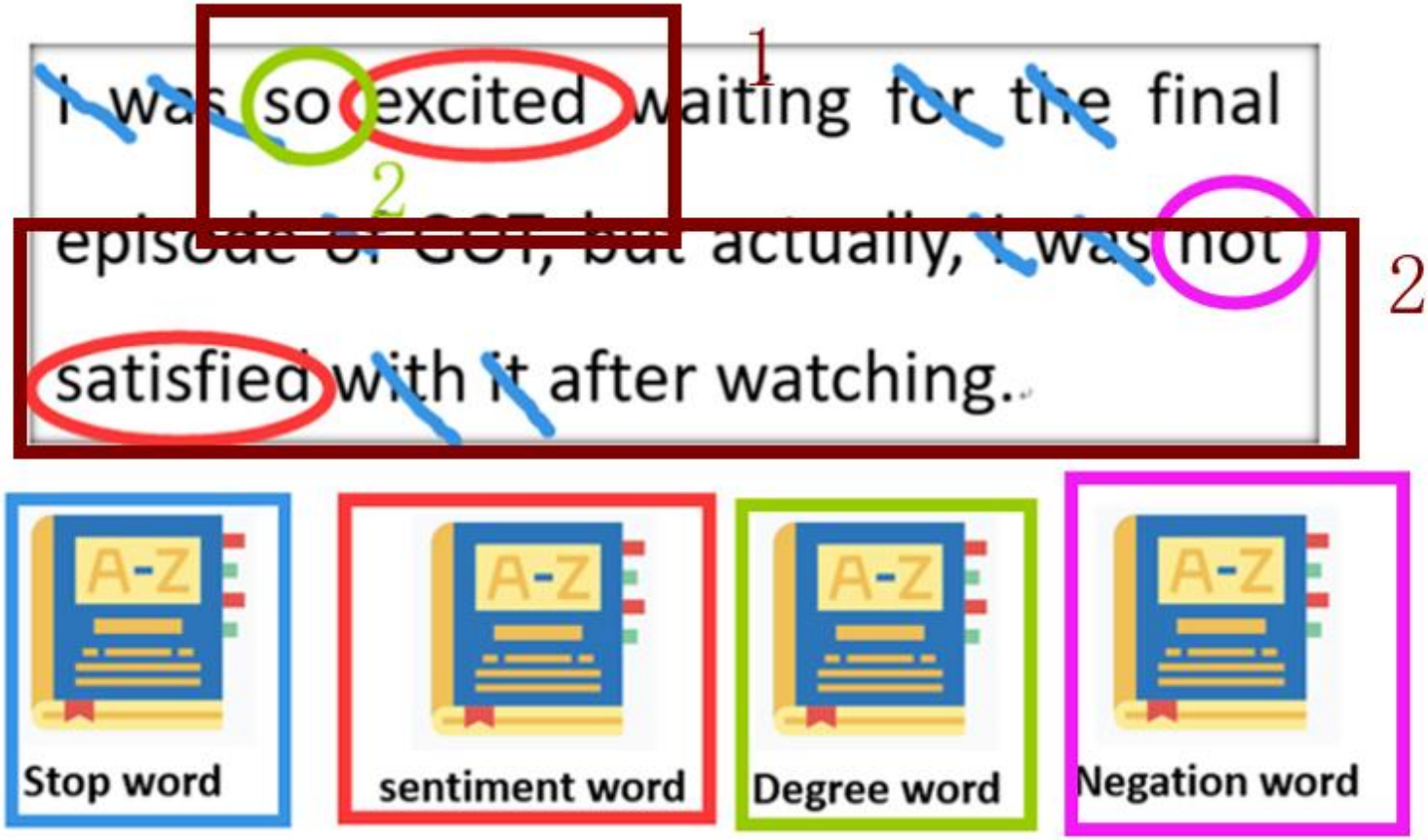
  "Jieba" (Chinese for "to stutter") Chinese text segmentation

- Stop word filtering

  Stop words are those words which are unlikely to assist further semantic understanding of computer algorithms. Such as *the*, *is*, *at*, *which*, and *on*.

Sentiment analysis: Lexicon-based method flow

Score = (2*1) + (-1*1) = 1



Sentiment analysis: Lexicon-based method flow

Lexicons description

Lexicon	Source
Sentiment lexicon	Dalian University of Technology
Degree word lexicon	HowNet
Negation word lexicon	Unofficial source (28 words)
Stop word lexicon	Harbin Institute of Technology

Sentiment analysis: Lexicon-based method flow

Algorithm evaluation

	Positive	Negative
Precision	0.63	0.74
Recall	0.29	0.11
F_1	0.40	0.19

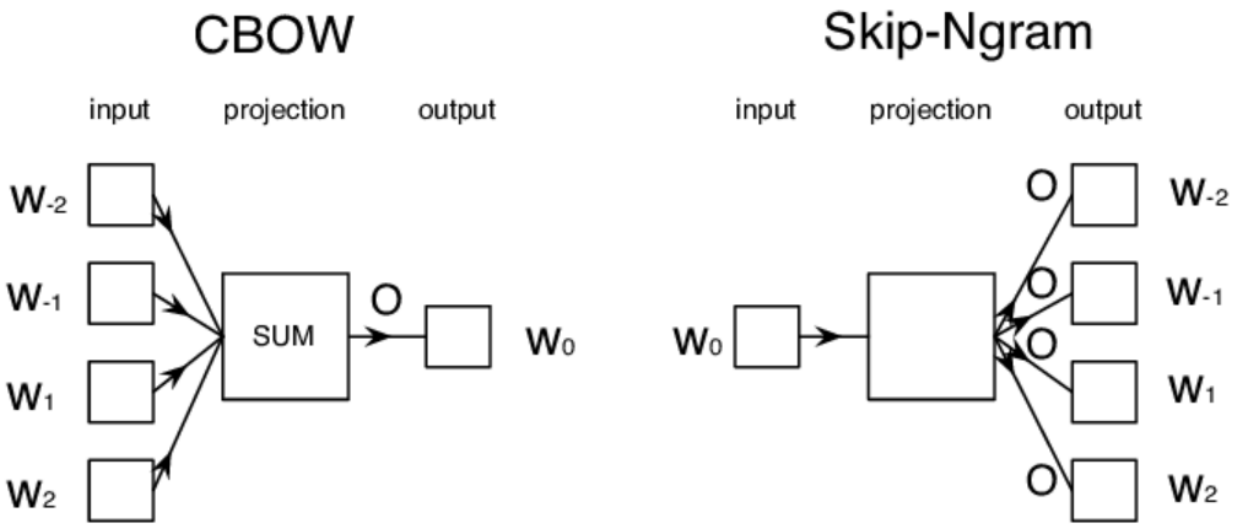
Sentiment analysis: Machine-learning-based method

- Vectorization
- Classification

Sentiment analysis: Machine-learning-based method

Vectorization

Word2Vec Model (Word embedding Model)– Google 2013



CBOW and Skip-Gram Architectures

Sentiment analysis: Machine-learning-based method

Algorithms and results

The optimal hyperparameters and the evaluations of classification methods

Methods	Parameters	Accuracy	F_1 score
Logistic Regression	C=0.1, Max_iter=100, Penalty='l2', solver='lbfgs'	0.73	0.73
K Nearest Neighbors	N_neighbors=5, Leaf_size=30, metric='minkowski'	0.68	0.68
Support Vector Machine (linear kernel)	C=1, kernel='linear'	0.69	0.69
Naïve Bayes (GaussianNB)	Var_smoothing= 1e-9	0.68	0.66
Random Forest	n_estimators=200, min_samples_split=800, max_depth=10	0.72	0.71

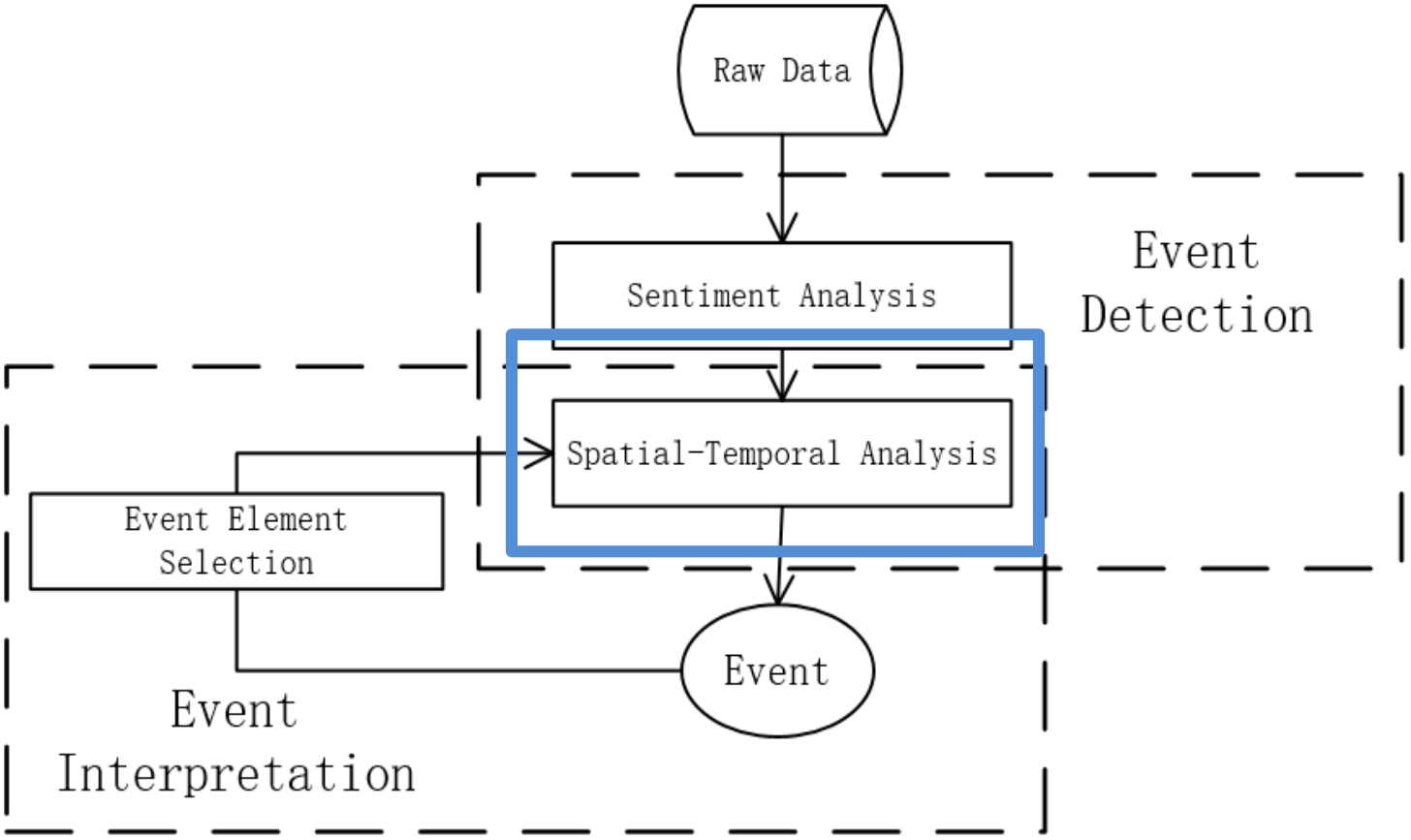
Sentiment analysis: Machine-learning-based method

Classification result - LG

Classification result

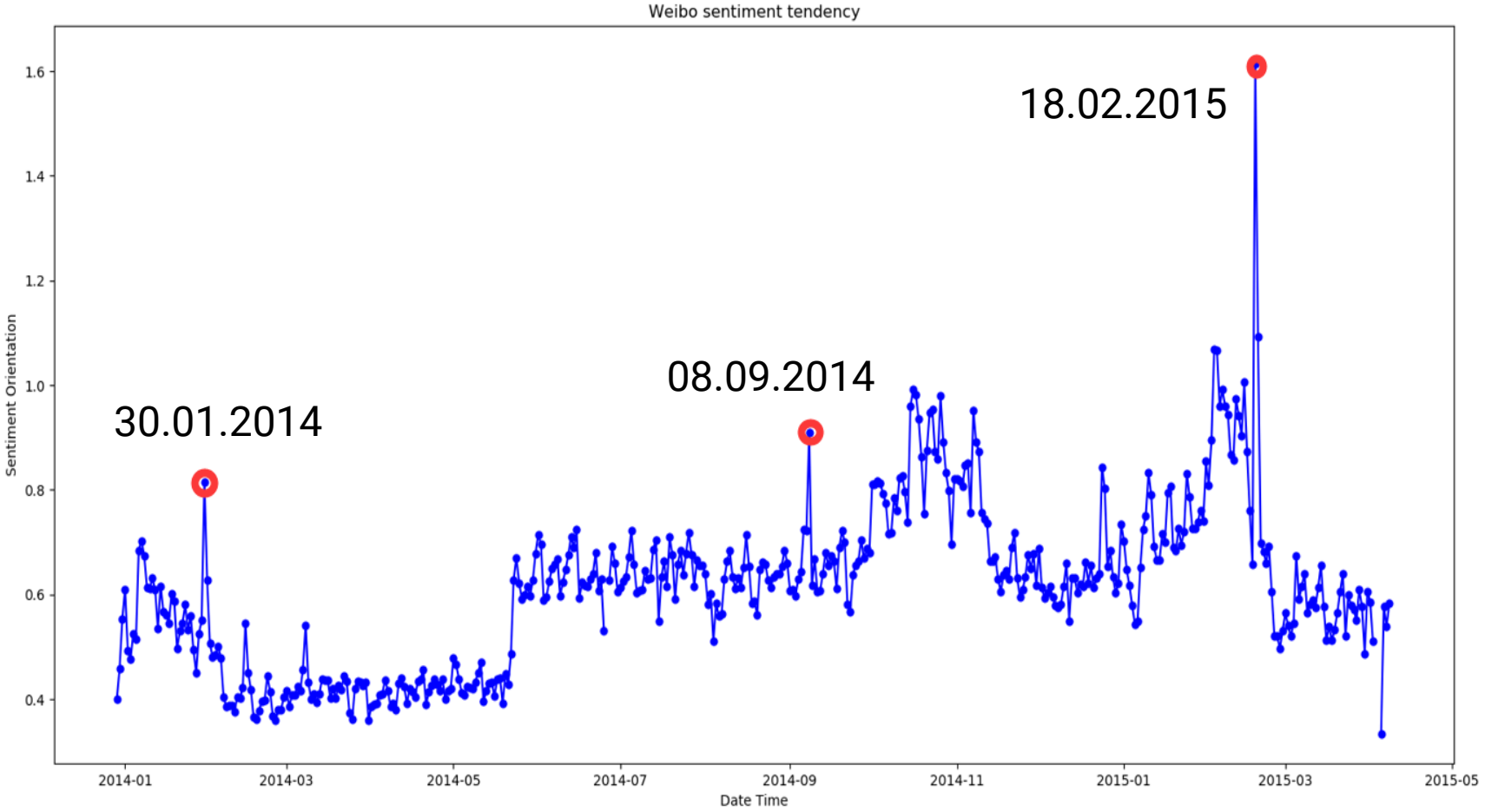
Classes	Quantity
Positive	3939058
Negative	6811234
No sentiment	196411
Total	10946703

Workflow


Raw Data
Sentiment Analysis
Event Detection
Spatial-Temporal Analysis
Event Element Selection
Event
Event Interpretation

2. Spatial-temporal analysis-based event detection

Spatial-temporal analysis: Time series analysis



Spatial-temporal analysis: Event extraction

Word Cloud

The Word Cloud of 18.02.2015, local maximum

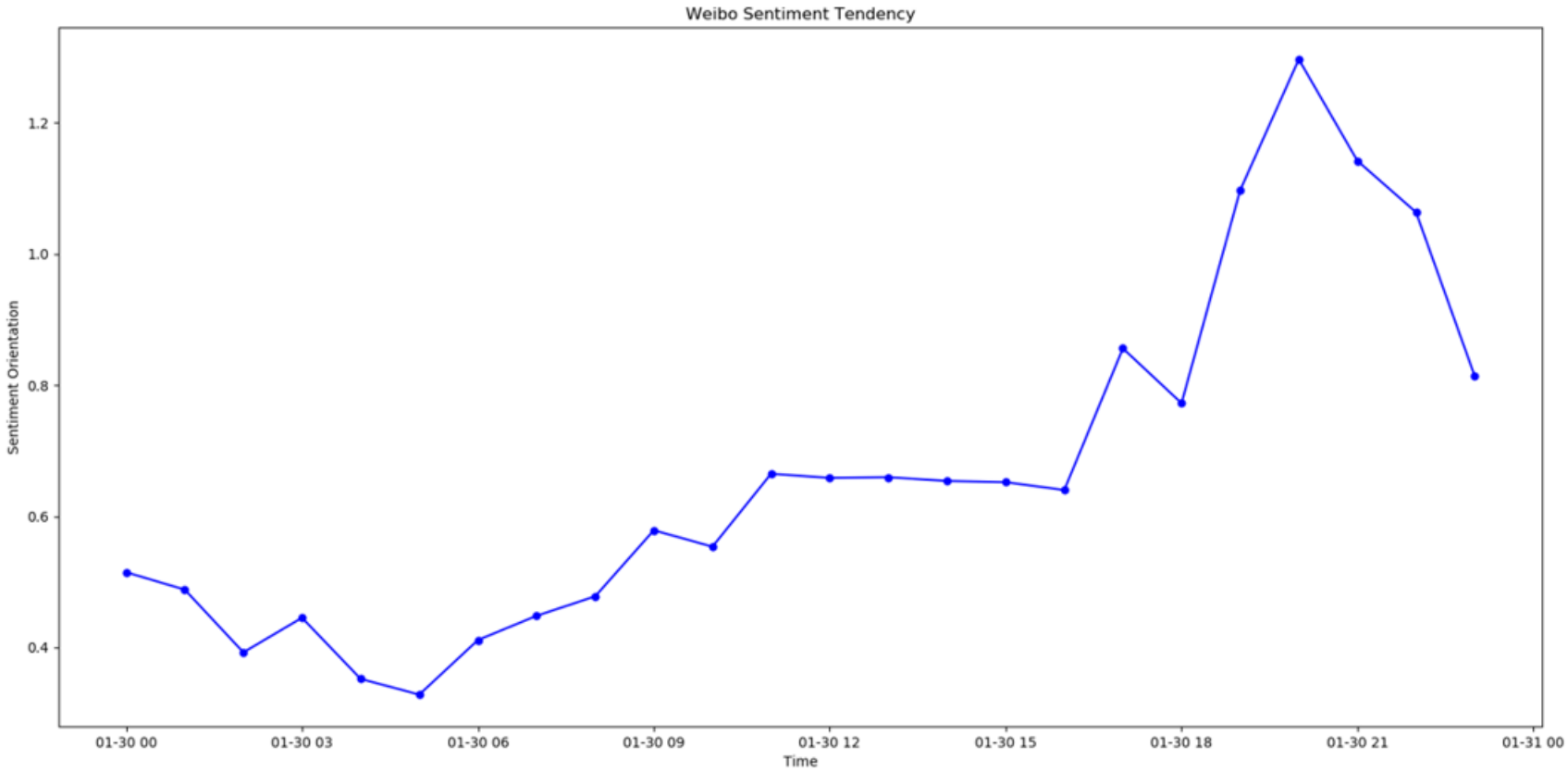
The Word Cloud of 08.09.2014, local maximum

The Word Cloud of 30.01.2014, local maximum

The keywords translation in Word Cloud

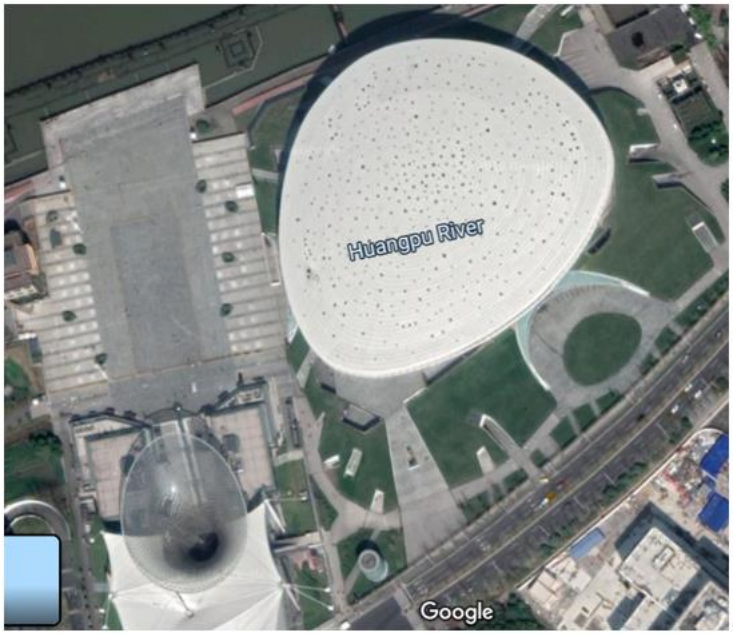
Date of maximum	Top frequent keywords
18.02.2015	"Spring Festival Gala", "Happy New Year", "Red Packets", "New Year's Eve", "Family", "New Year", "Family Dinner", "Red Lantern", "Good Wish", "Friends", etc.
08.09.2014	"Mid-Autumn", "Mid-Autumn Festival", "Moon", "Moon Cake", "Happy Holiday", "Enjoying the Moon", "Reunion", "Happy", "Blessing", etc.
30.01.2014	"Spring Festival Gala", "Happy New Year", "Spring Festival", "New Year's Eve", "Red Packets", "Expecting", "Friends", "Good Health", "Hope", etc.

Spatial-temporal Analysis: Small temporal scale event detection



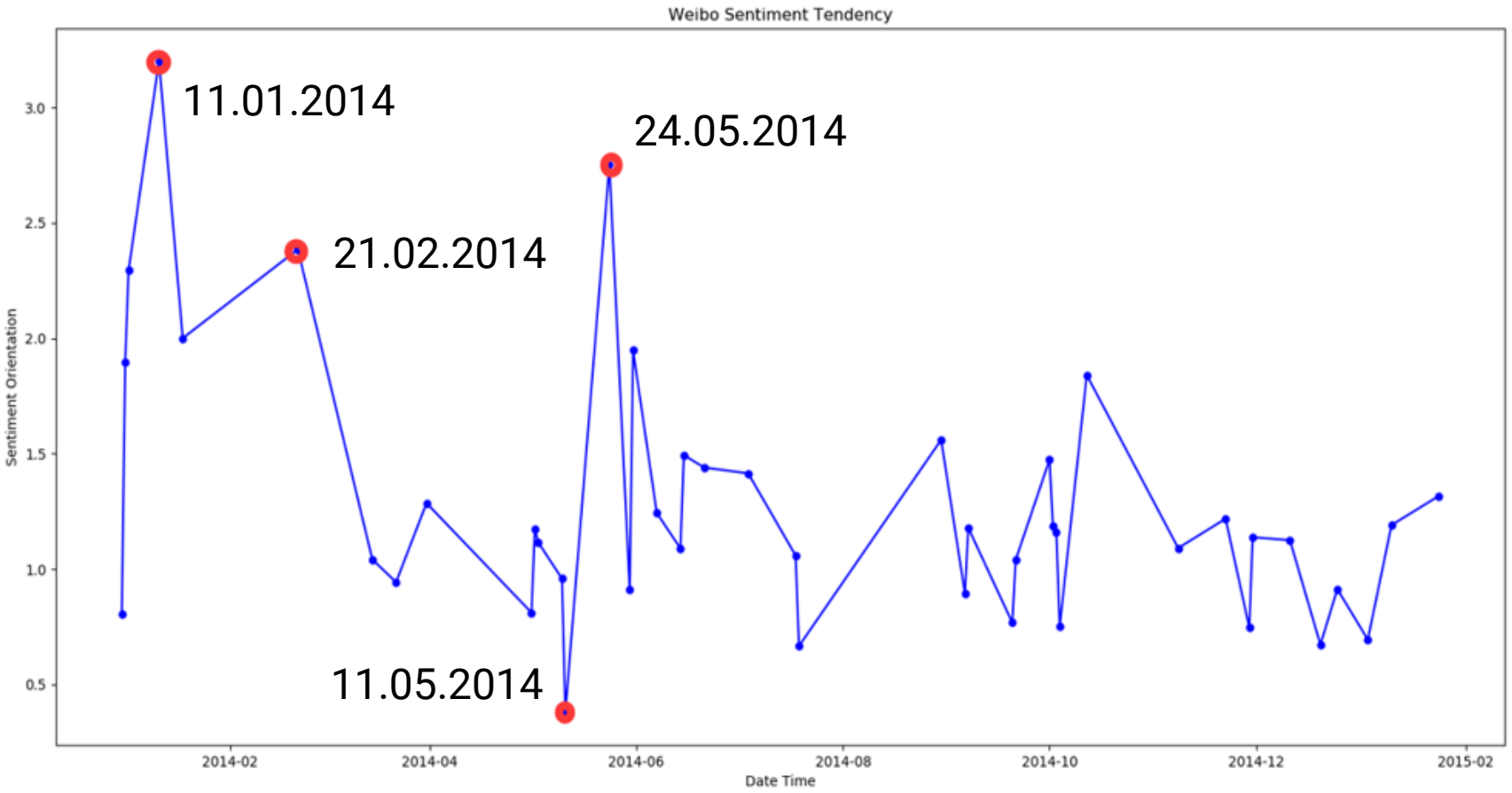
Time series analysis of the New Year's Eve of 2014

Spatial-temporal analysis: Small spatial scale local event detection



Mercedes-Benz Arena, Shanghai on Google map

Spatial-temporal Analysis: Small spatial scale local event detection



PSO time series plot of Mercedes-Benz Arena

Spatial-temporal analysis: Small spatial scale local event detection

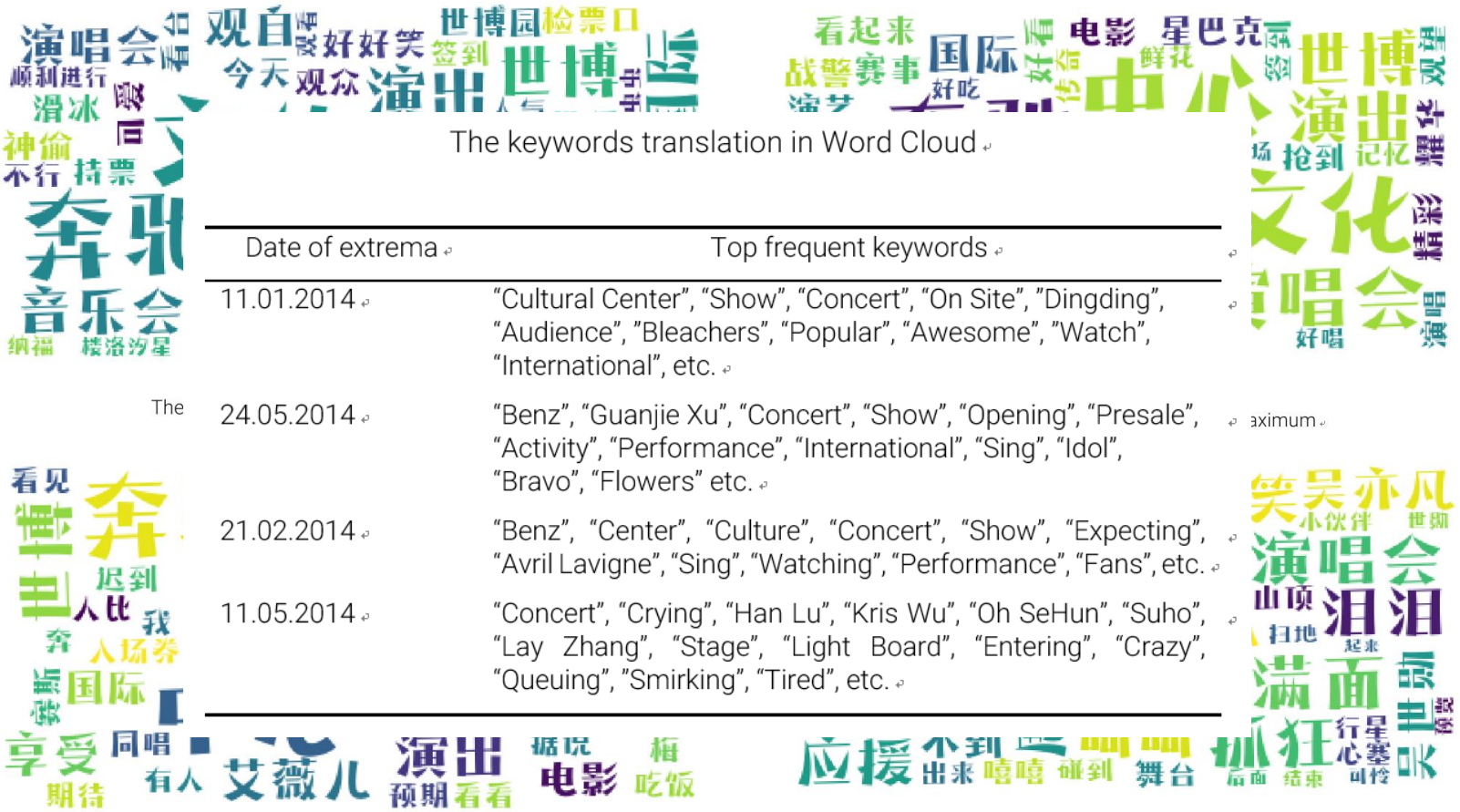
The Word Cloud of 11.01.2014, local maximum

The Word Cloud of 24.05.2014, local maximum

The Word Cloud of 21.02.2014, local maximum

The Word Cloud of 11.05.2014, local minimum

Spatial-temporal analysis: Small spatial scale local event detection



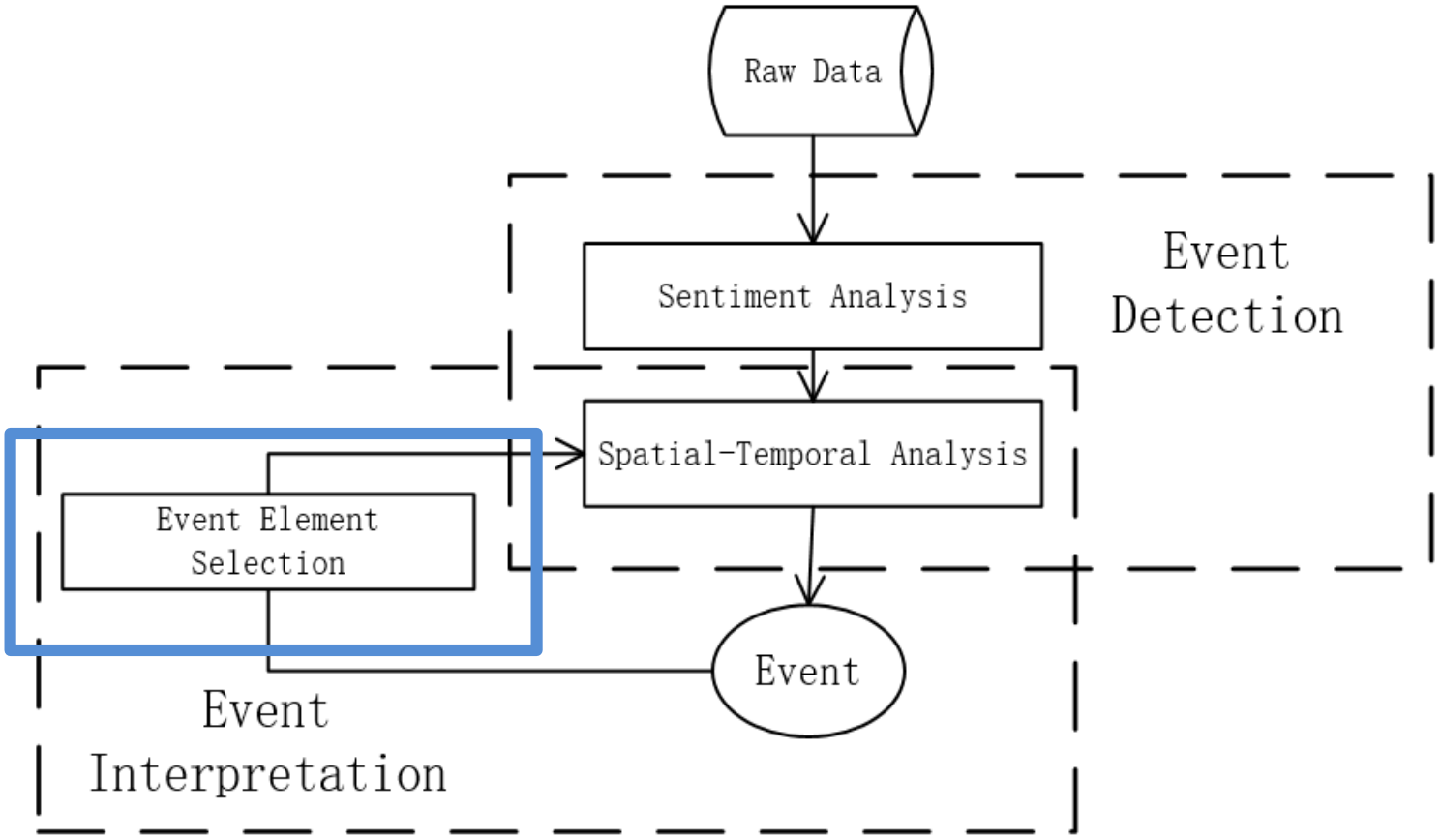
The keywords translation in Word Cloud

Date of extrema	Top frequent keywords
11.01.2014	"Cultural Center", "Show", "Concert", "On Site", "Dingding", "Audience", "Bleachers", "Popular", "Awesome", "Watch", "International", etc.
24.05.2014	"Benz", "Guanjie Xu", "Concert", "Show", "Opening", "Presale", "Activity", "Performance", "International", "Sing", "Idol", "Bravo", "Flowers" etc.
21.02.2014	"Benz", "Center", "Culture", "Concert", "Show", "Expecting", "Avril Lavigne", "Sing", "Watching", "Performance", "Fans", etc.
11.05.2014	"Concert", "Crying", "Han Lu", "Kris Wu", "Oh SeHun", "Suho", "Lay Zhang", "Stage", "Light Board", "Entering", "Crazy", "Queuing", "Smirking", "Tired", etc.

The Word Cloud of 21.02.2014, local maximum

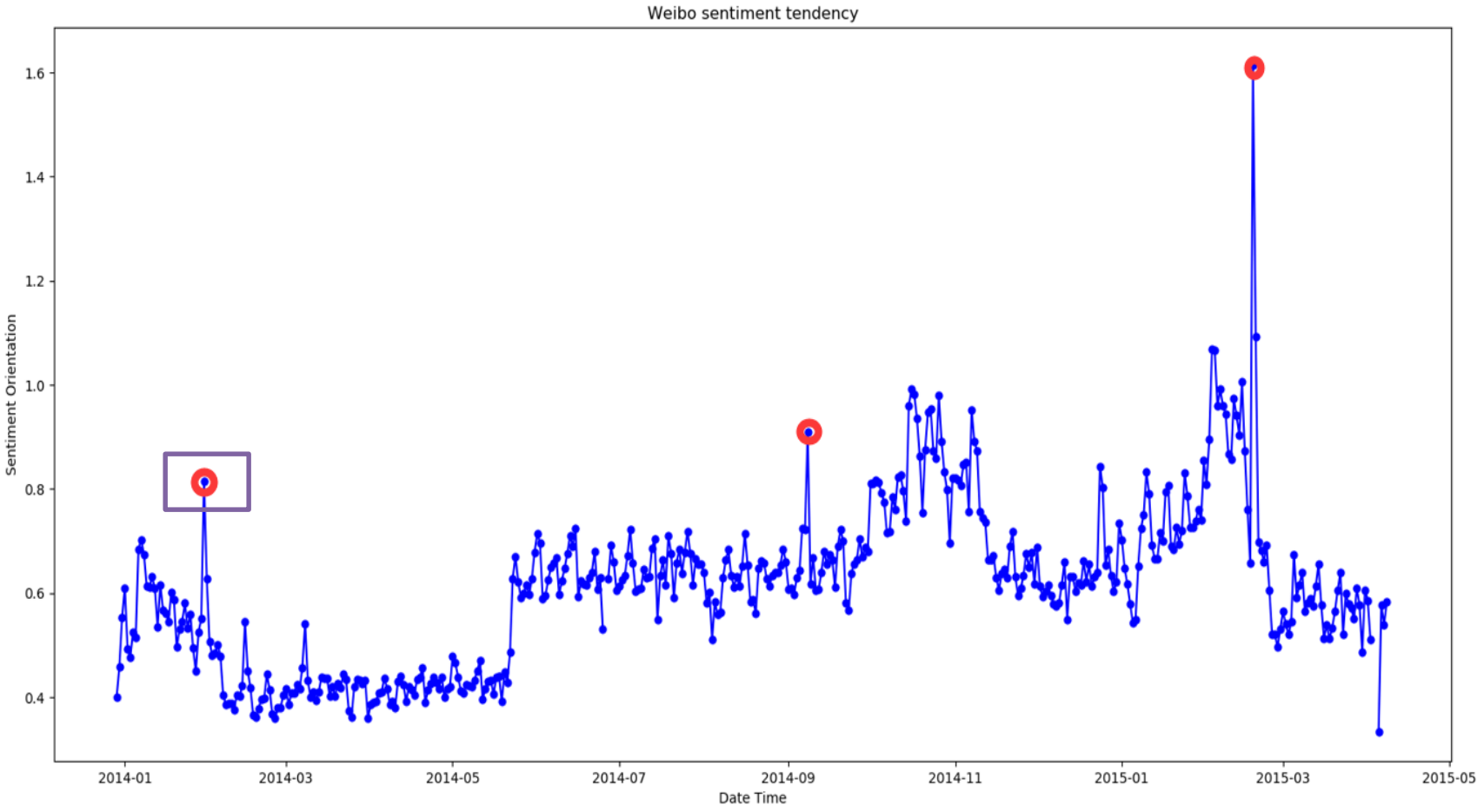
The Word Cloud of 11.05.2014, local minimum

Workflow

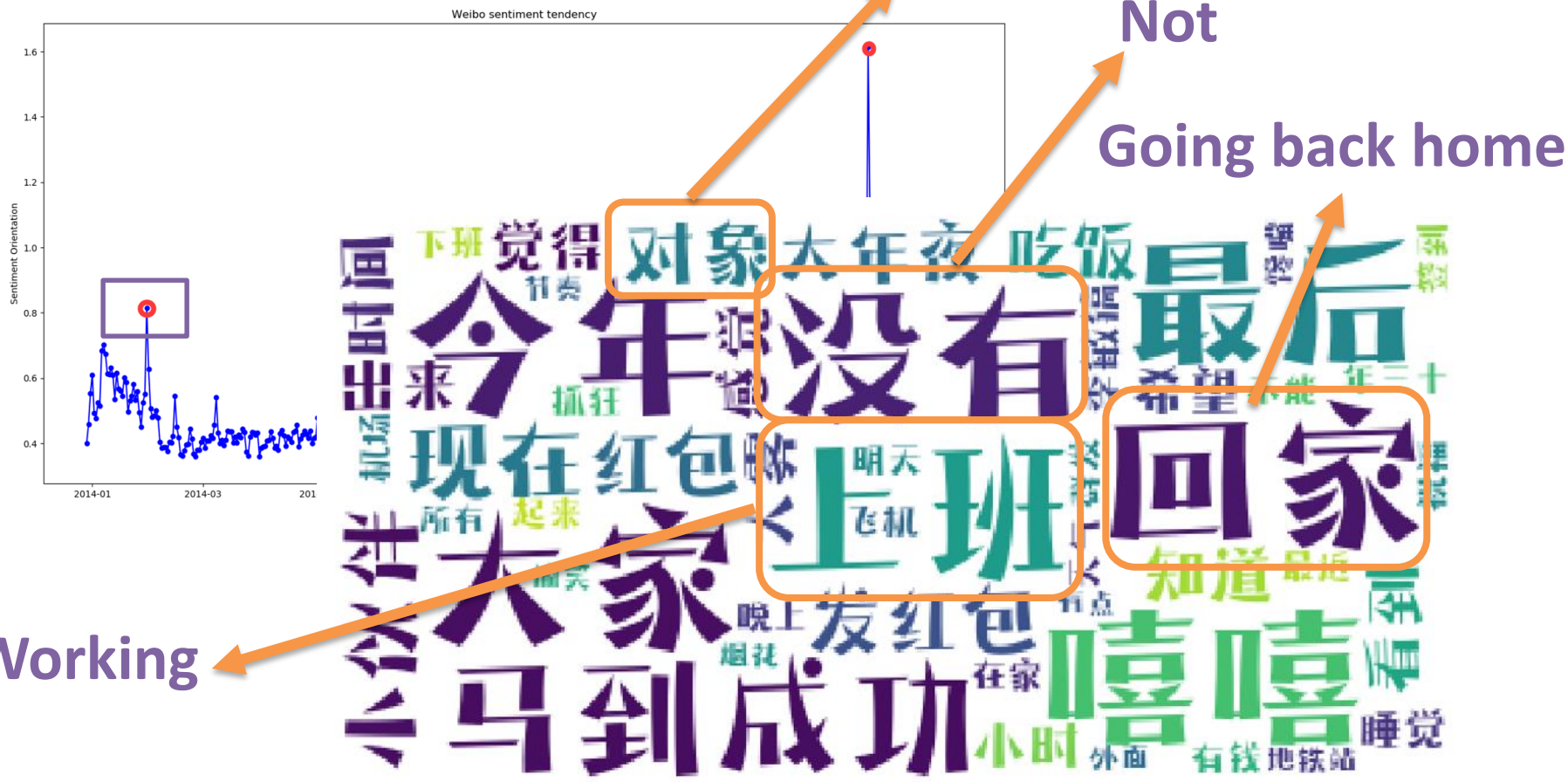

Raw Data
Sentiment Analysis
Event Detection
Spatial-Temporal Analysis
Event Element Selection
Event
Event Interpretation

3. Spatial-temporal analysis-based event interpretation

Spatial-temporal analysis: Event interpretation

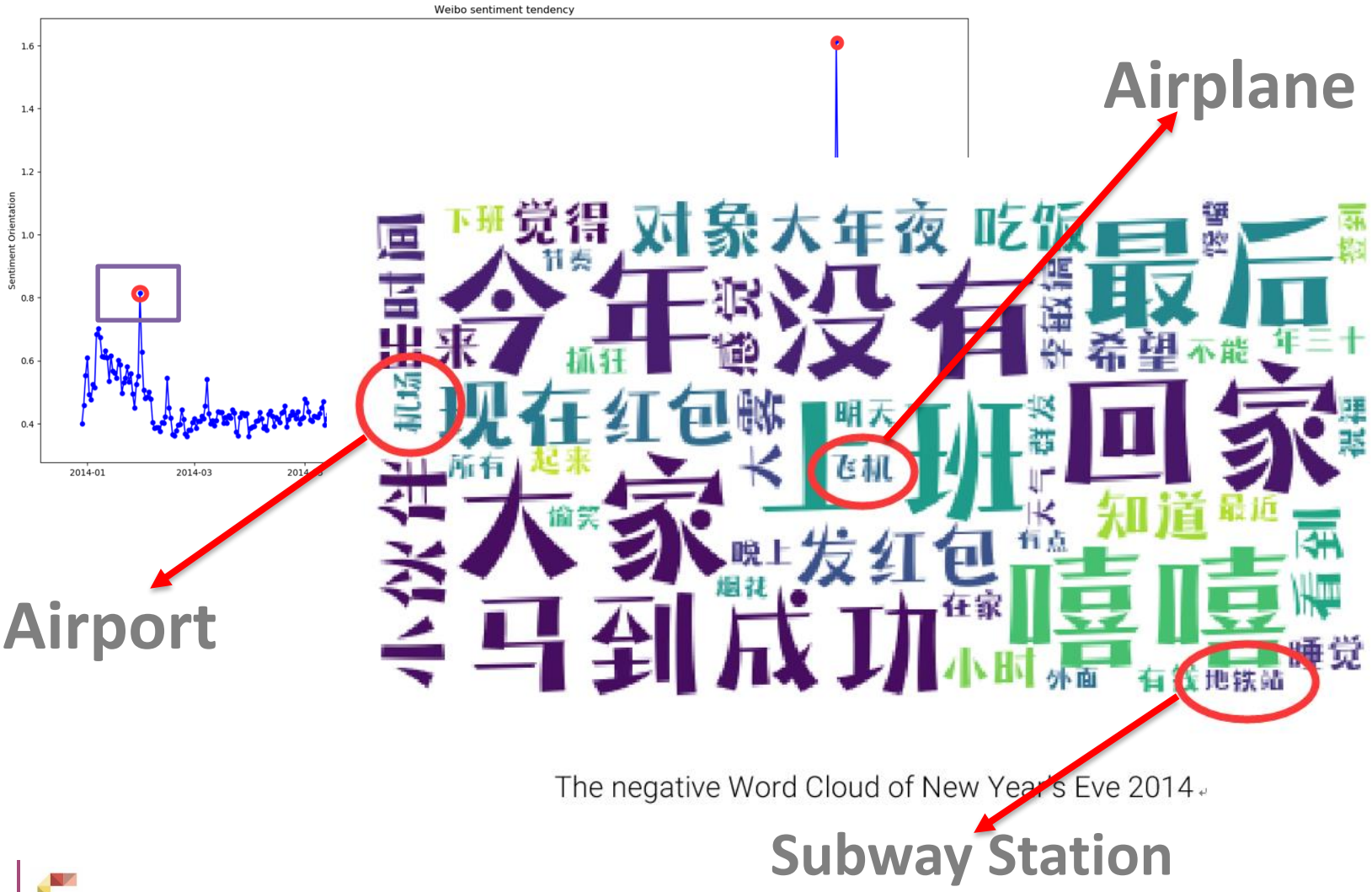


Spatial-temporal analysis: Event interpretation



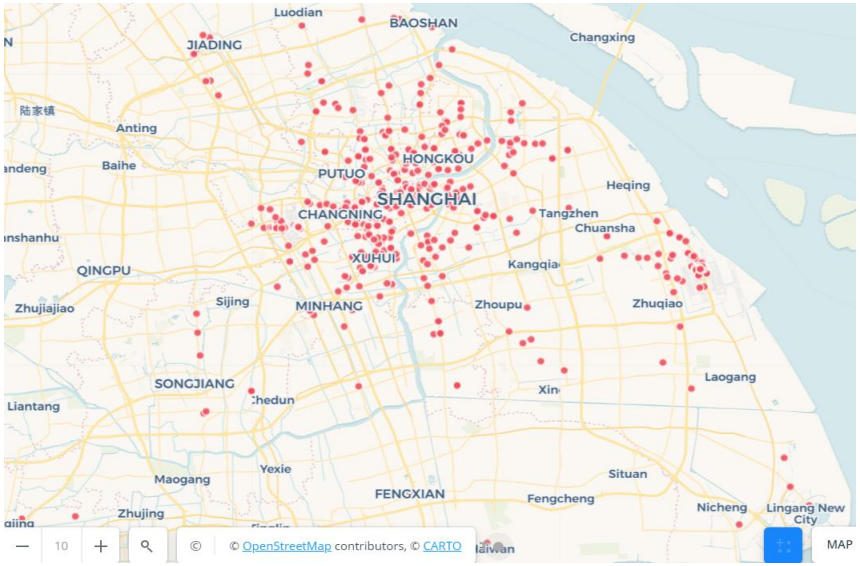
The negative Word Cloud of New Year's Eve 2014

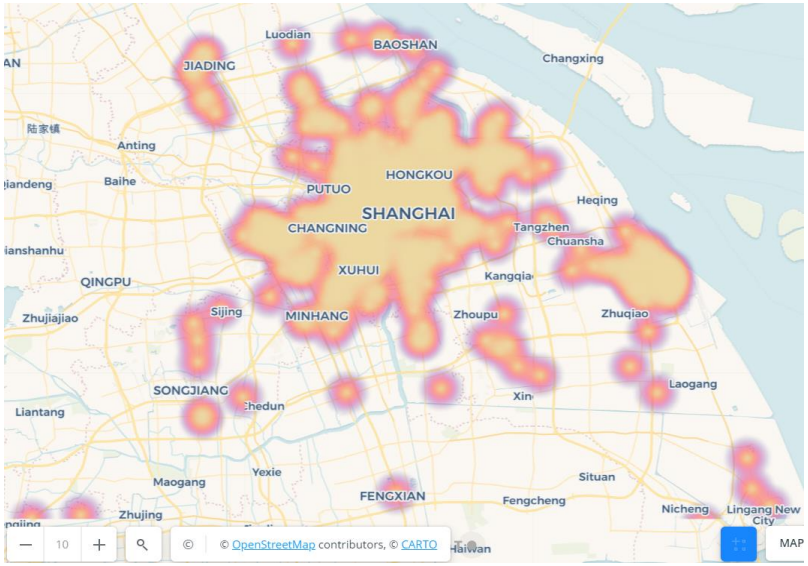
Spatial-temporal analysis: Event interpretation



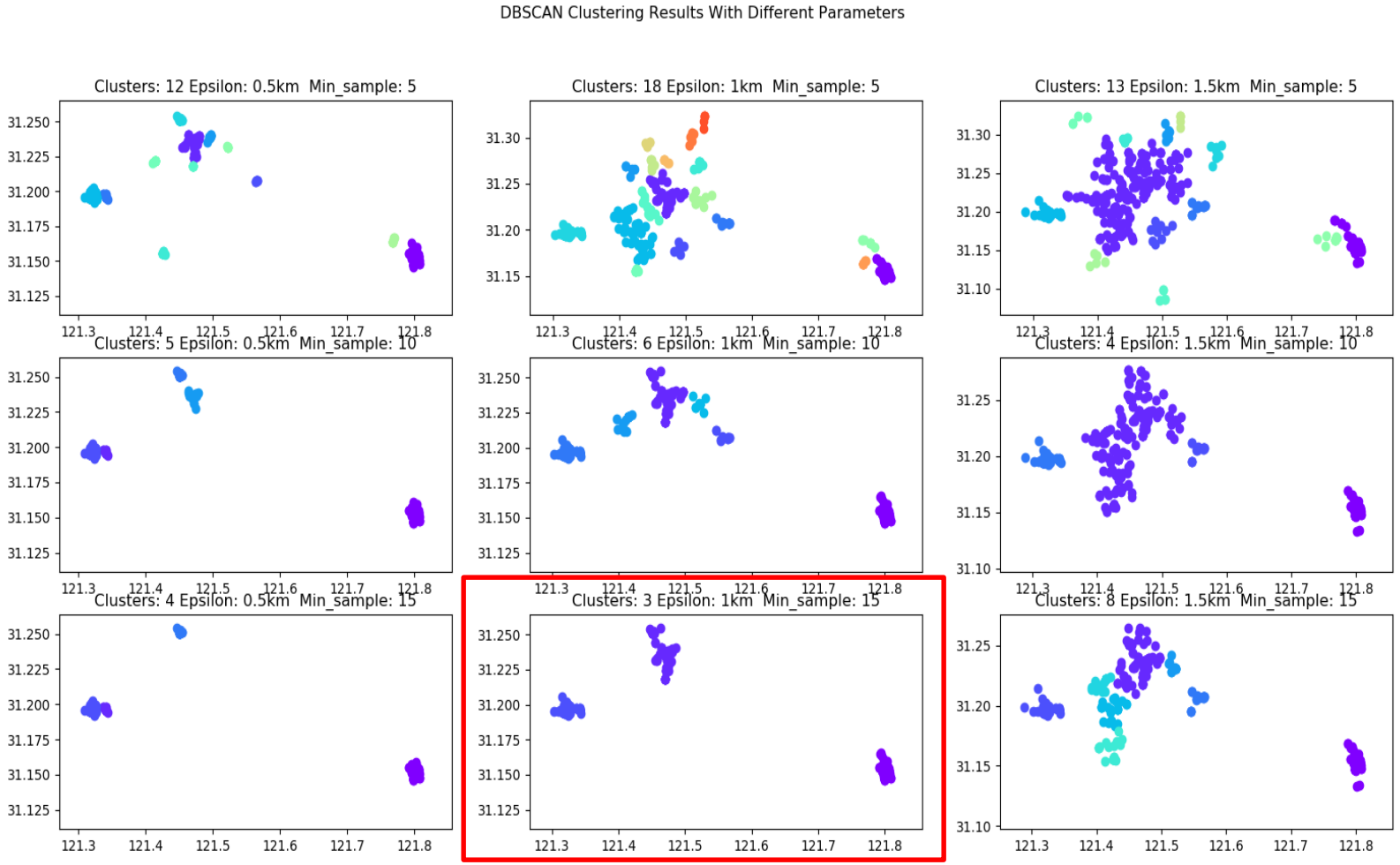
The negative Word Cloud of New Year's Eve 2014

Spatial-temporal analysis: Spatial clustering analysis

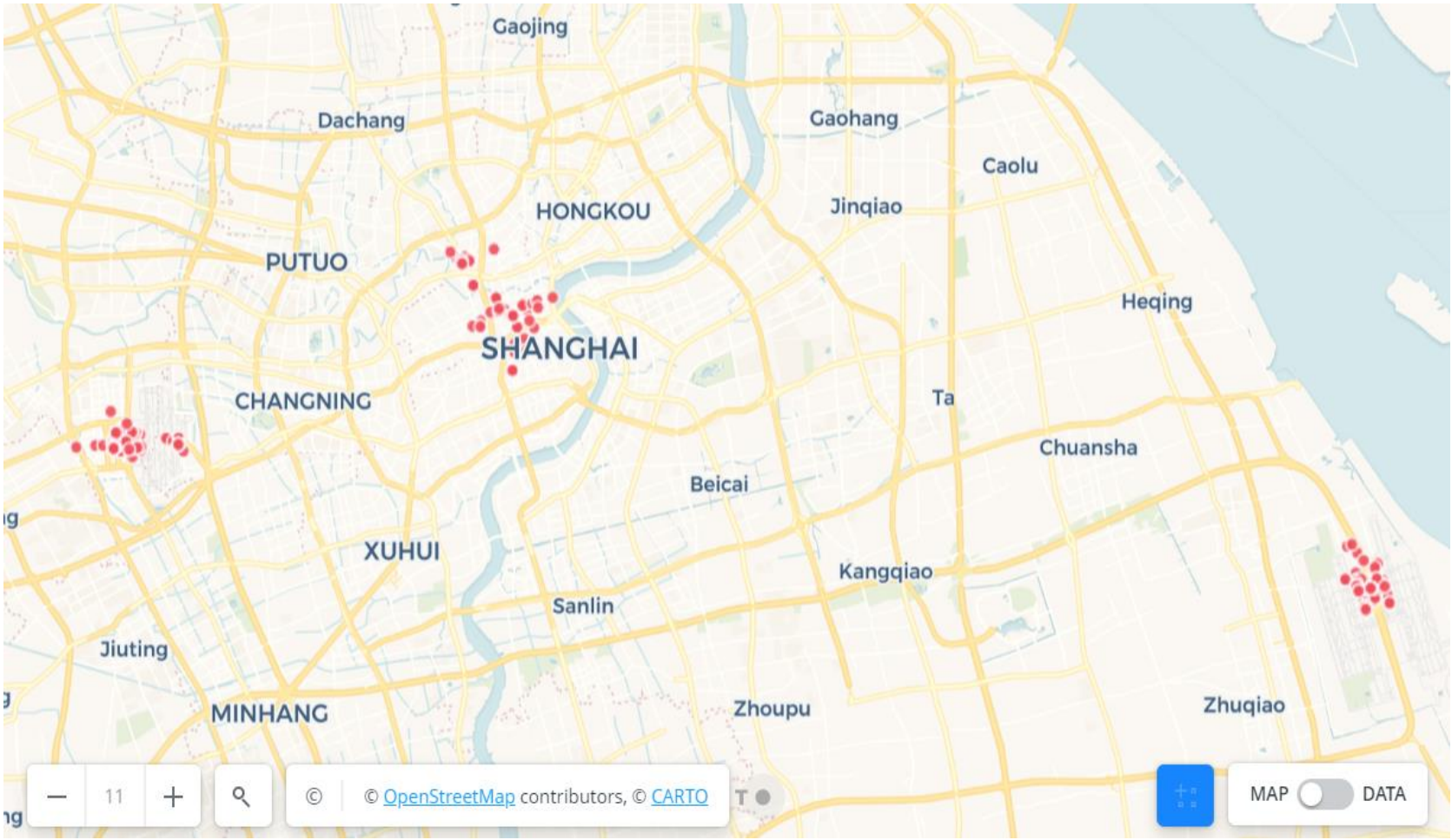




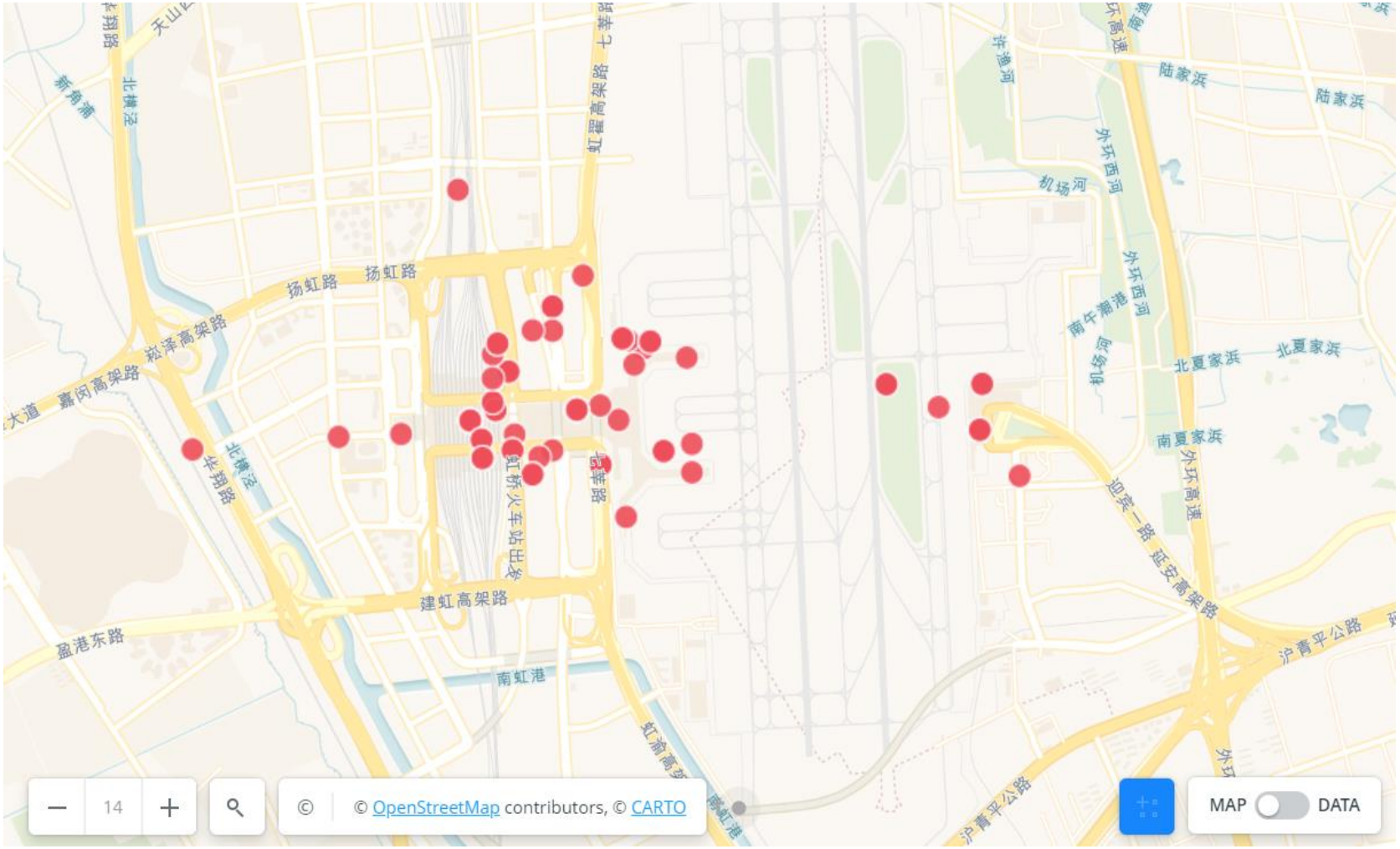
Spatial-temporal analysis: Spatial clustering analysis



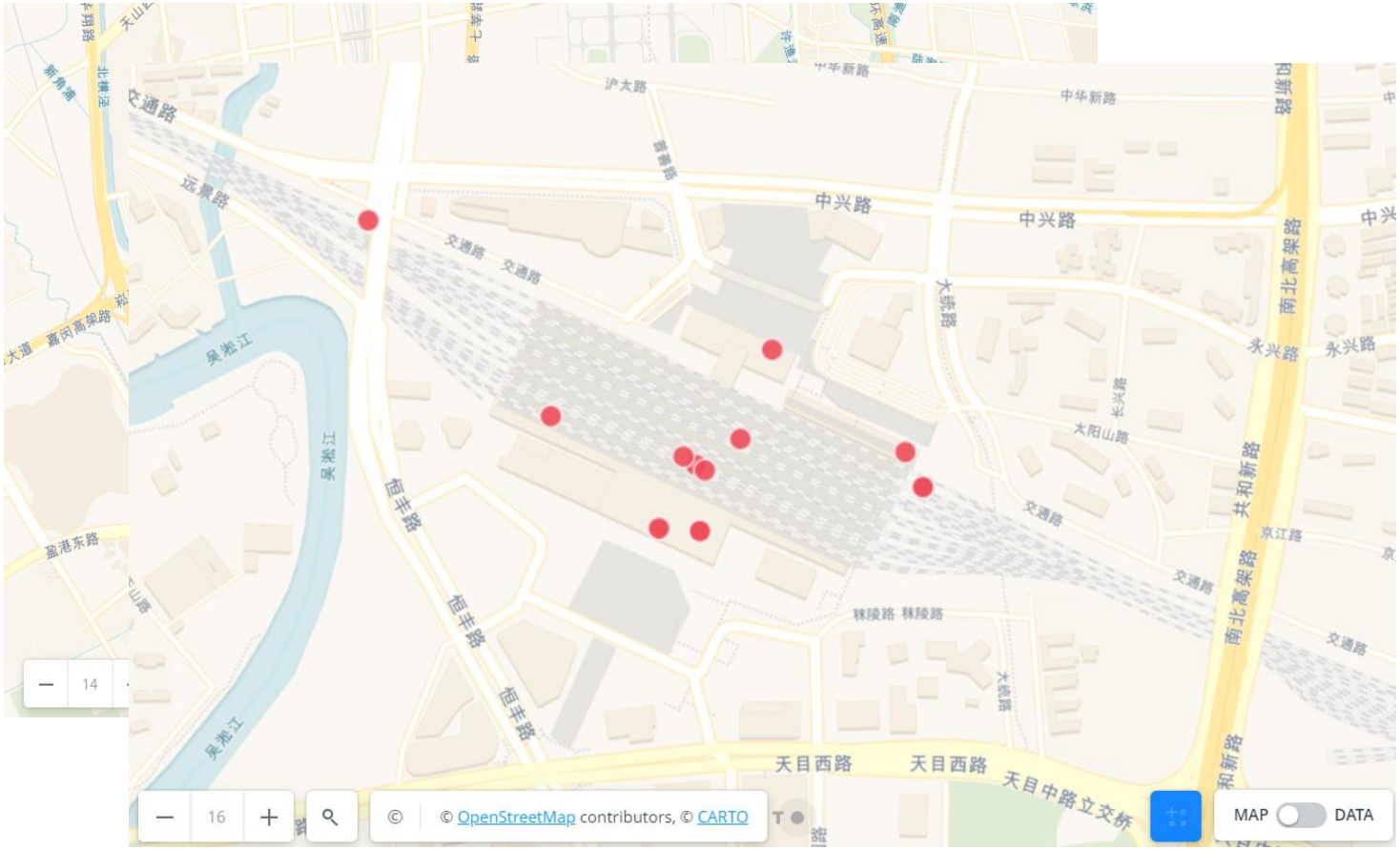
Spatial-temporal analysis: Spatial clustering analysis



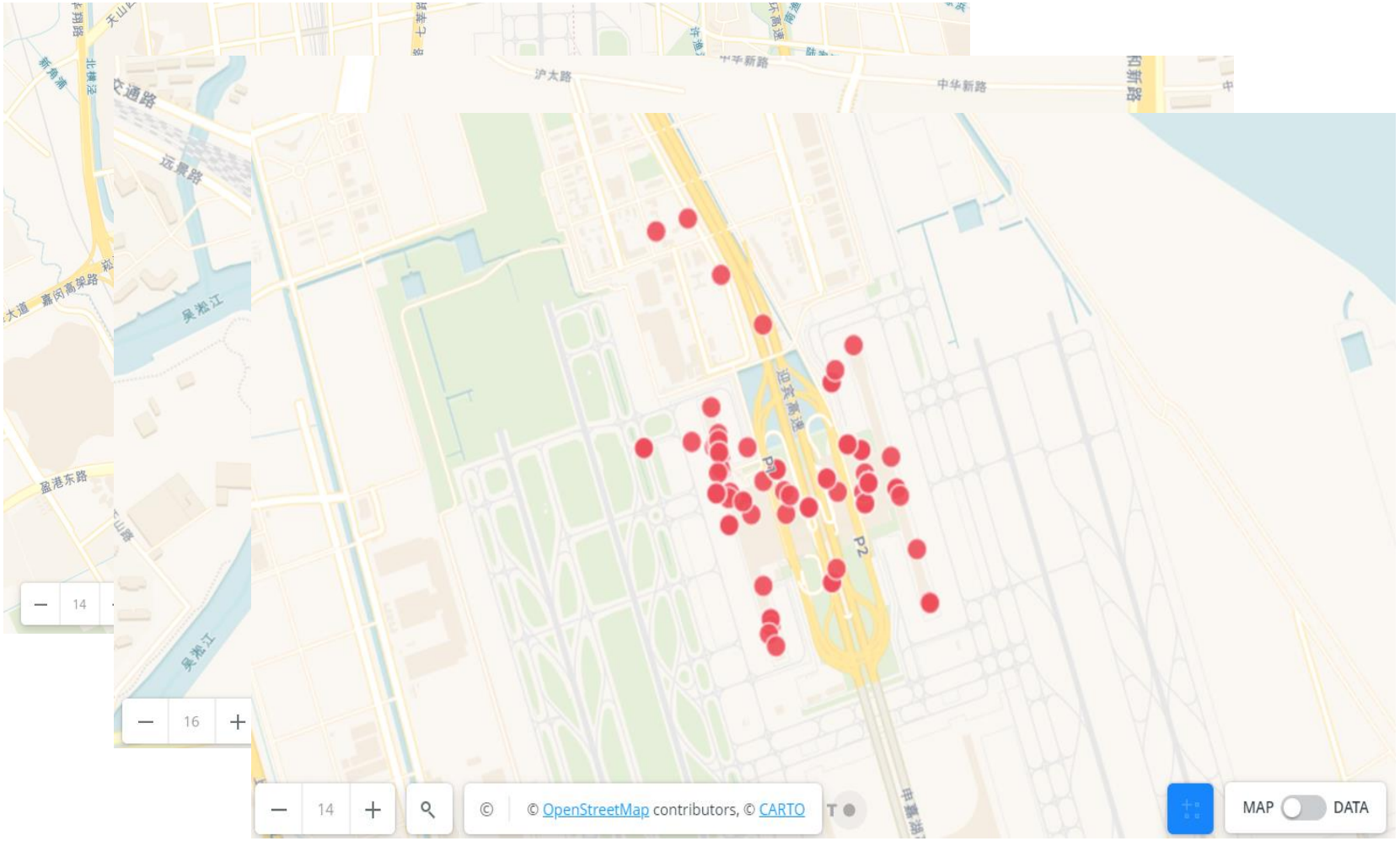
Spatial-temporal analysis: Spatial clustering analysis



Spatial-temporal analysis: Spatial clustering analysis



Spatial-temporal analysis: Spatial clustering analysis



Conclusion

- Sentiment analysis
- Population sentiment orientation (PSO)
- Multi-scale
- Event interpretation
- Wider range

Outlook

- Sentiment analysis
- An interactive web event analysis system
- Event Monitoring and Early Warning System
- Multi-source data fusion

“Part of the journey is the end”

and thank you all

Lin Che

Munich, 24.09.2019