Visualizing and Analyzing Location Based Social Media Data – a case study of EU Migration Crisis

Sagnik Mukherjee

Dirk Burghardt, Eva Hauthal (TUD)

Francisco Porras Bernardez (TUW)

TECHNISCHE UNIVERSITÄT DRESDEN

Cartography M.Sc.

Layout

- Introduction

- Objectives

- Background

- Objective 1

- Objective 2

- Objective 3

- Conclusion

Layout

- Introduction

- Objectives

- Background

- Objective 1

- Objective 2

- Objective 3

- Conclusion

3

Introduction

- 170,000 geo-tagged tweets from 2016 till 2021, filtered using relevant hashtags

- Tweets are in English, German, Dutch, Spanish, Italian and French

- PostgreSQL server on virtual machine to store raw and HyperLogLog data

- Python 3.9 for analysis and exploration : Temporal, Spatial and Topical facets according to Dunkel et al. (2019)

- Python 3.9, ~~QGIS and ArcGIS Pro~~ for visualization

Layout

- Introduction

- **Objectives**

- Background

- Objective 1

- Objective 2

- Objective 3

- Conclusion

Objectives

- How can the HyperLogLog data format be used for exploring spatio-temporal patterns?

- How to visualize the results concisely and clearly?

- How does public reactions vary across geographic space and time?

Layout

- Introduction

- Objectives

- **Background**

- Objective 1

- Objective 2

- Objective 3

- Conclusion

Background: HyperLogLog(HLL)

- Probabilistic algorithm to count cardinality

- Forms the basis to develop the privacy aware data structure

- Allows quantitative estimations with error rates of 2% (Flajolet P. et al. 2007)

Background: Facets of LBSN

- Used to characterize reactions of an event (Dunkel et. al 2019)

- Spatial : The *where* of the post

- Temporal : The *when* of the post

- Social: The *who* of the post

- Topical: The *what* of the post

Background: Typicality (Hauthal et. al., 2021)

Typicality of top 6 hashtags
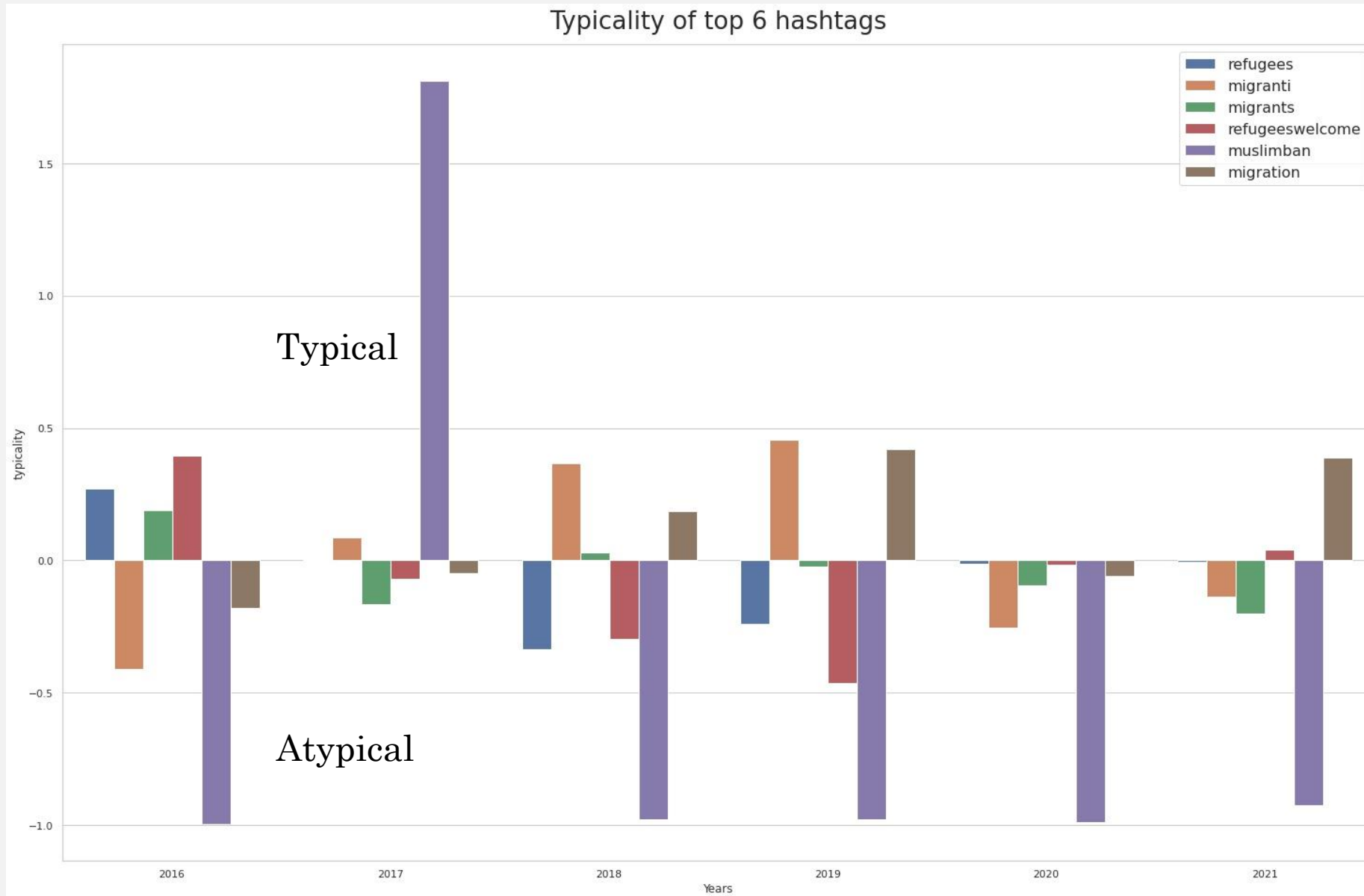
Background: Opinion Analysis with Social Media Data

- Hauthal et al. (2019) used hashtags for opinion analysis on the Brexit

  referendum

- This work however has a different case study

Background: Visualizations

- Kumar et al. (2014) suggests using trendlines, word clouds and heat maps for twitter data visualization

- Interactive visualizations used to mitigate contextual information loss (Heer et al.2008)

- They also improve cognition for complex datasets (Dix and Ellis 1998)

Layout

- Introduction

- Objectives

- Background

- **Objective 1**

- Objective 2

- Objective 3

- Conclusion

13

Layout

- Introduction

- Objectives

- Background

- **Objective 1**

- Objective 2

- Objective 3

- Conclusion

14

Objective 1 : HLL Workflow

```
┌──────────────────┐          ┌──────────────────┐
│                  │          │                  │
│    Raw Data      │─────────▶│    Processing    │
│                  │          │                  │
└──────────────────┘          └──────────────────┘
                                       │
                                       │
                                       ▼
┌──────────────────┐          ┌──────────────────┐
│                  │          │                  │
│ Jupyter Notebooks│◀─────────│   PGSQL Server   │
│                  │          │                  │
└──────────────────┘          └──────────────────┘
```

Objective 1 : HLL

HLL Twitter Data

Normal Twitter Data

```
Settings   Help

HLL_auto.ipynb

Code                                    Python 3

TAG_A ='refugees'
TAG_B = 'refugeeswelcome'
TAG_C = 'migrants'
COLUMN = "user_hll"

three_df(TAG_A,TAG_B,TAG_C,COLUMN)
```
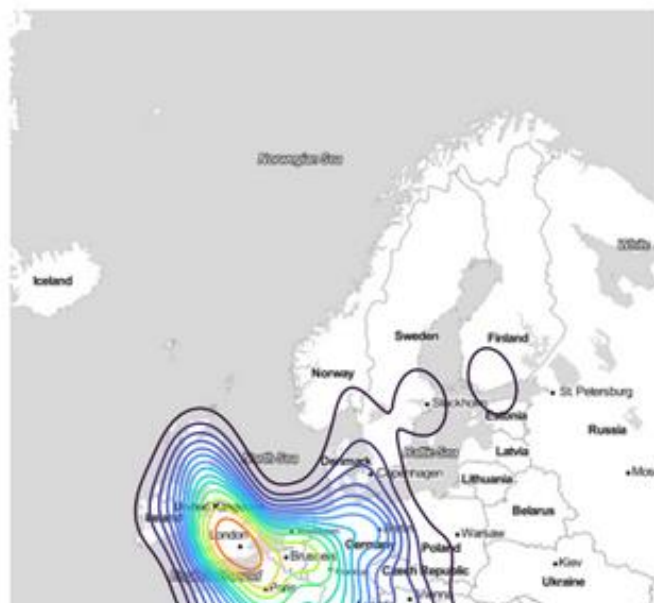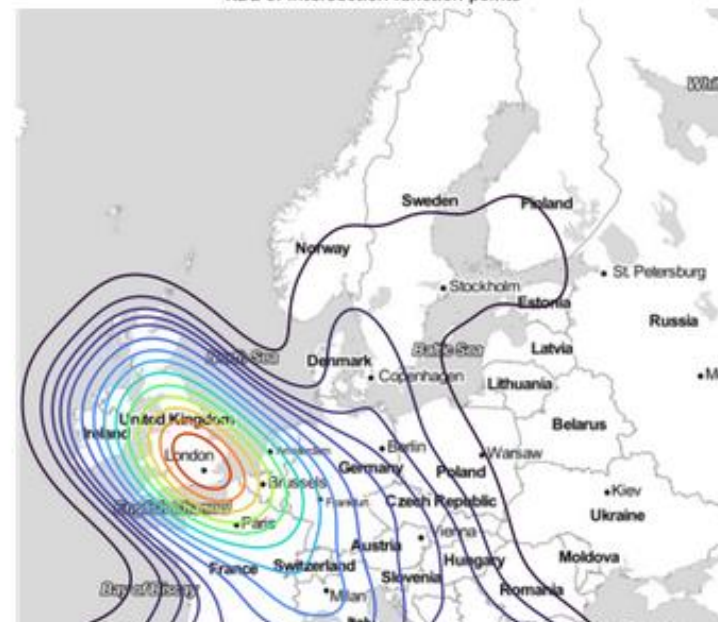
```
6784 distinct posts used REFUGEES
5743 distinct posts used REFUGEESWELCOME
3907 distinct posts used MIGRANTS
9627 distinct total posts which had either refugees or migrants (union)
11151 distinct total posts which had either refugees or refugeeswelcome (union)
9130 distinct total posts which had either refugeeswelcome or migrants (union)
1376 distinct posts with hashtags with refugees and refugeeswelcome (intersection)
1064 distinct posts with hashtags with refugees and migrants (intersection)
520 distinct posts with hashtags with refugeeswelcome and migrants (intersection)
Union Count : 13666 Intersection Count : 192
CPU times: user 54.2 s, sys: 231 ms, total: 54.4 s
Wall time: 1min
```
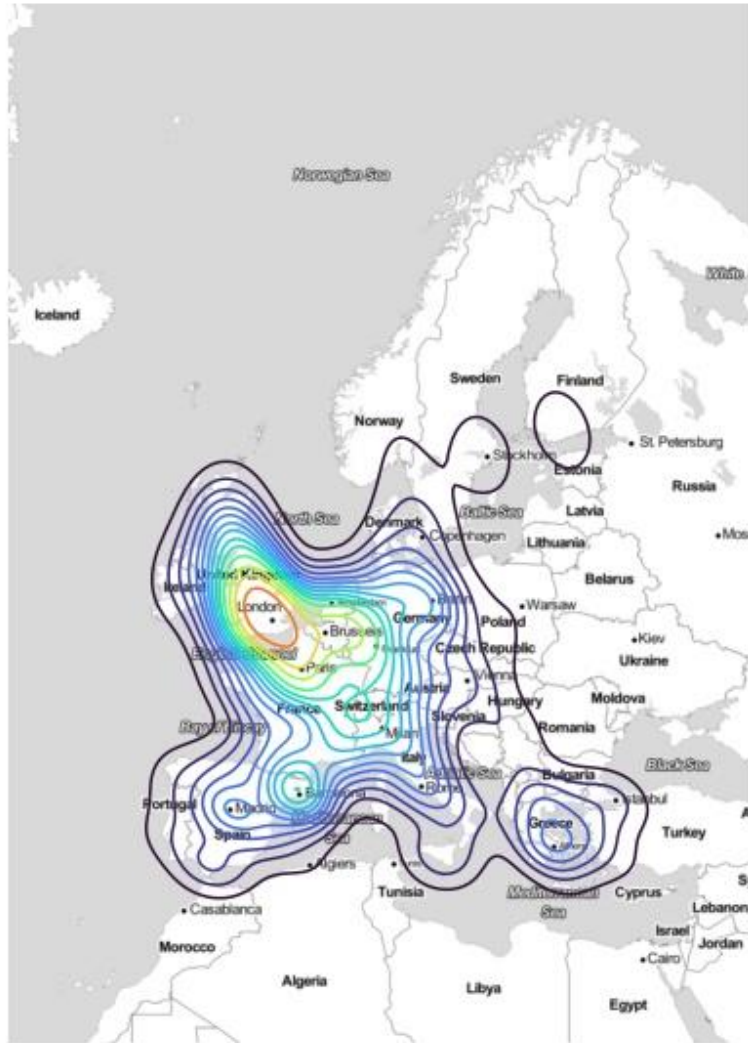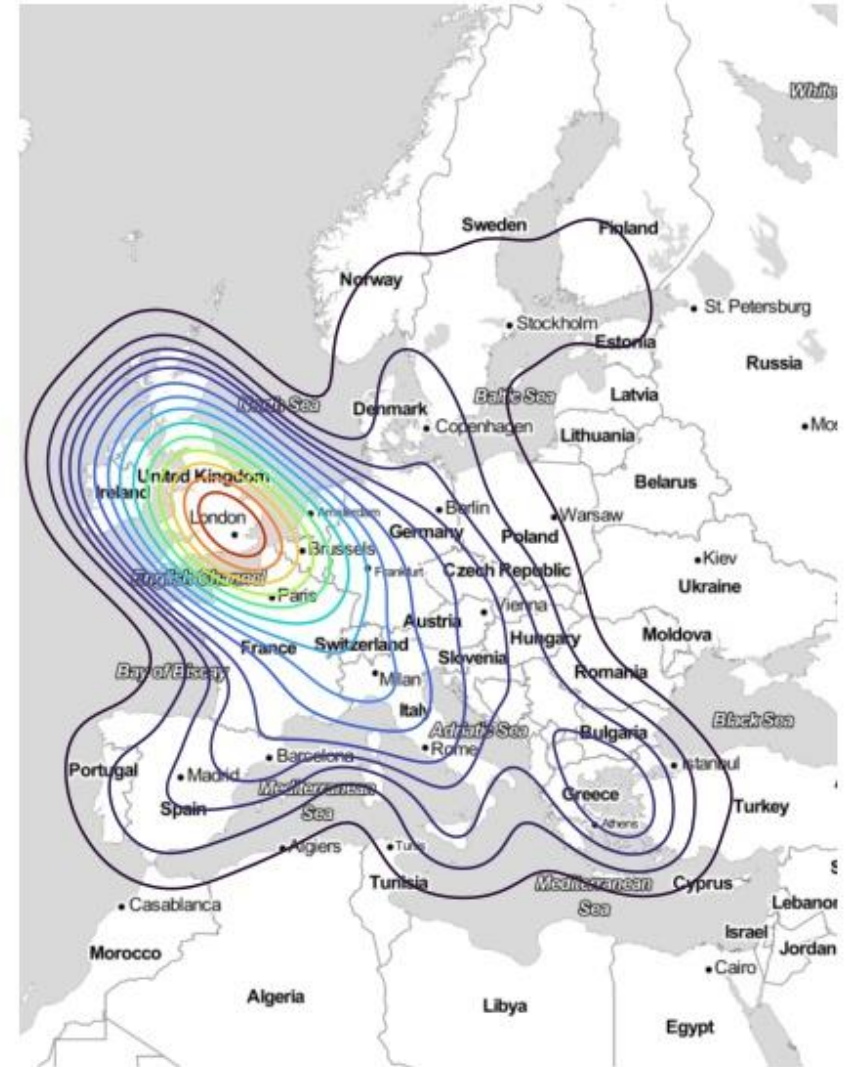
KDE of union function points

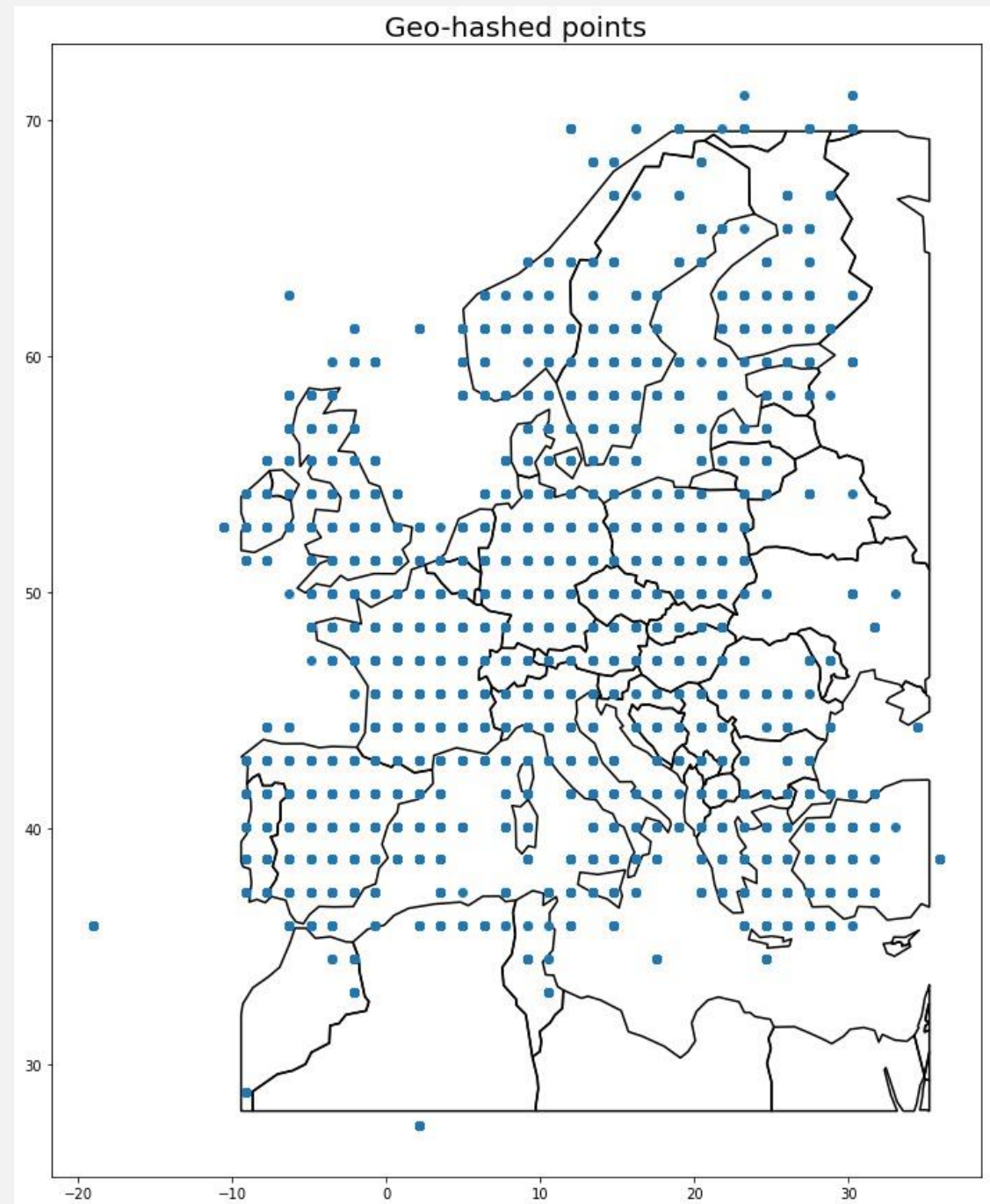KDE of intersection function points

Idle   Saving completed   Mode: Command   Ln 8, Col 35   HLL_auto.ipynb

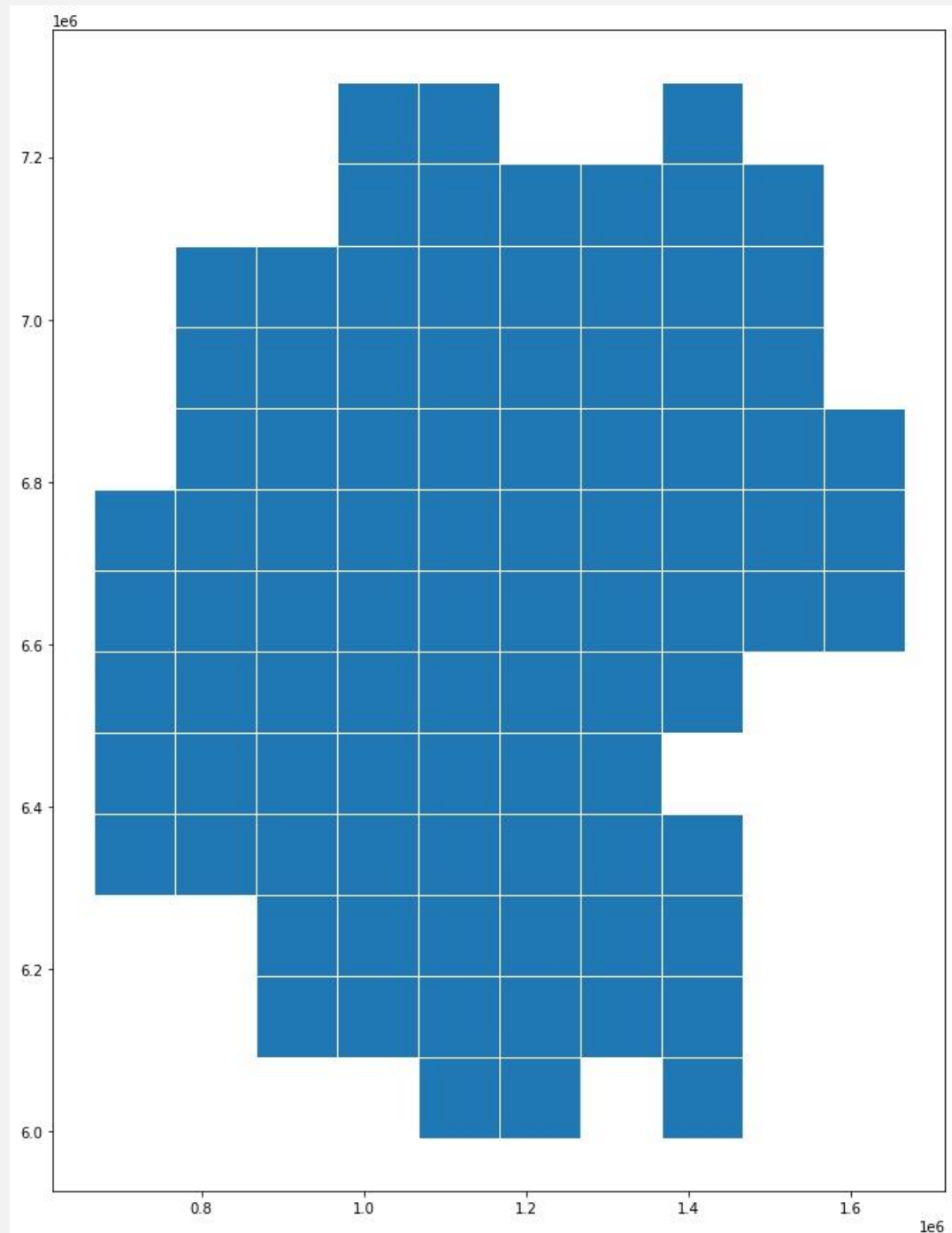USER_HLL – Union and Intersection

Geo-hashing

- Space is subdivided into two halves over and over until required precision is reached

- Used for storing large number of co-ordinates
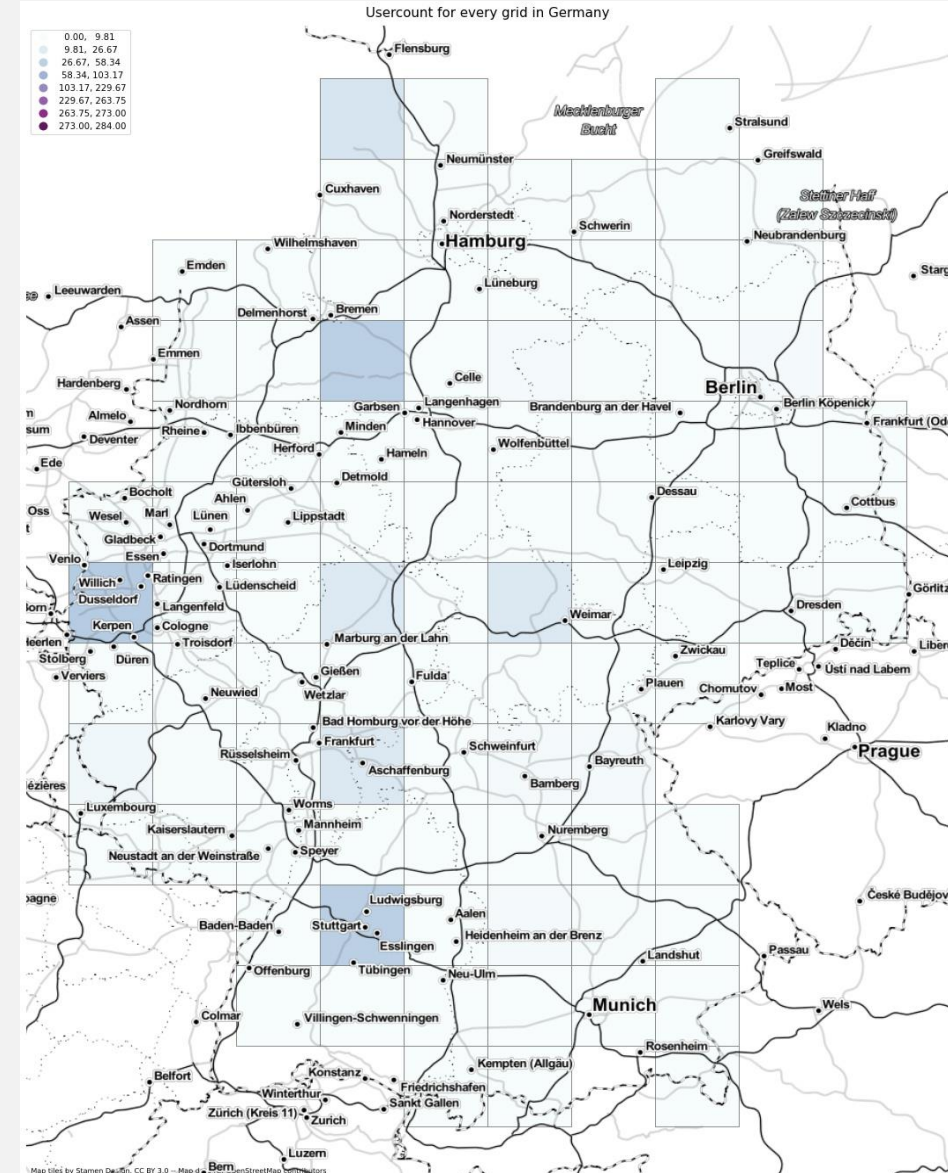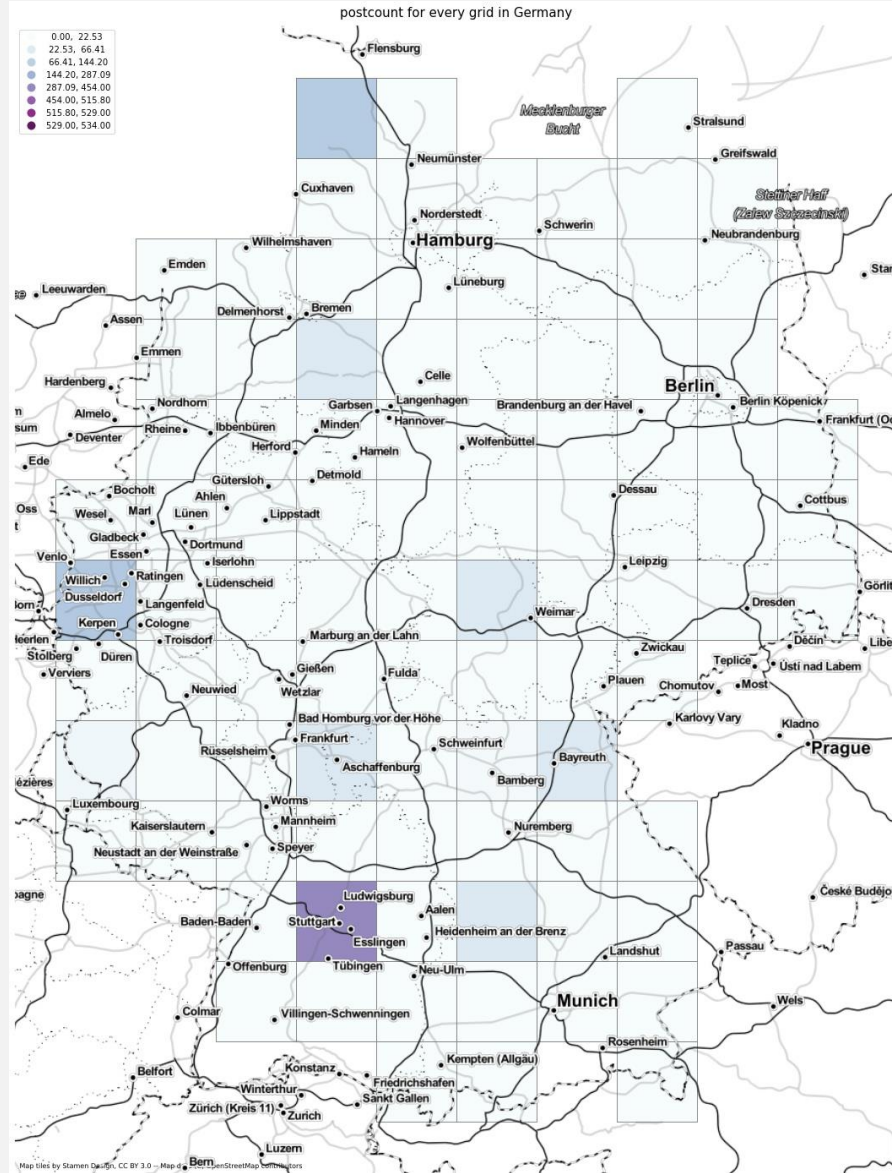
- The lack of precision can be used to protect privacy

Geo-hashed points

Geo-hashing : Grids

- Grids of 250 km x 250 km size

- Geo-hashed points are assigned to grids.

- Acts as an HLL alternative to KDE

Geo-hashed Grids Post Count and User Count

postcount for every grid in Germany

	0.00, 22.53
	22.53, 66.41
	66.41, 144.20
	144.20, 287.09
	287.09, 454.00
	454.00, 515.80
	515.80, 529.00
	529.00, 534.00

Usercount for every grid in Germany

|	
	0.00, 9.81
	9.81, 26.67
	26.67, 58.34
	58.34, 103.17
	103.17, 229.67
	229.67, 263.75
	263.75, 273.00
	273.00, 284.00

Objective 1 :Final Thoughts

- Robust at safeguarding the privacy of the users

- Very reliable quantitative estimations

- Flexibility for qualitative analysis with union and intersection functions

- Cumbersome to setup and work with (steep learning curve)

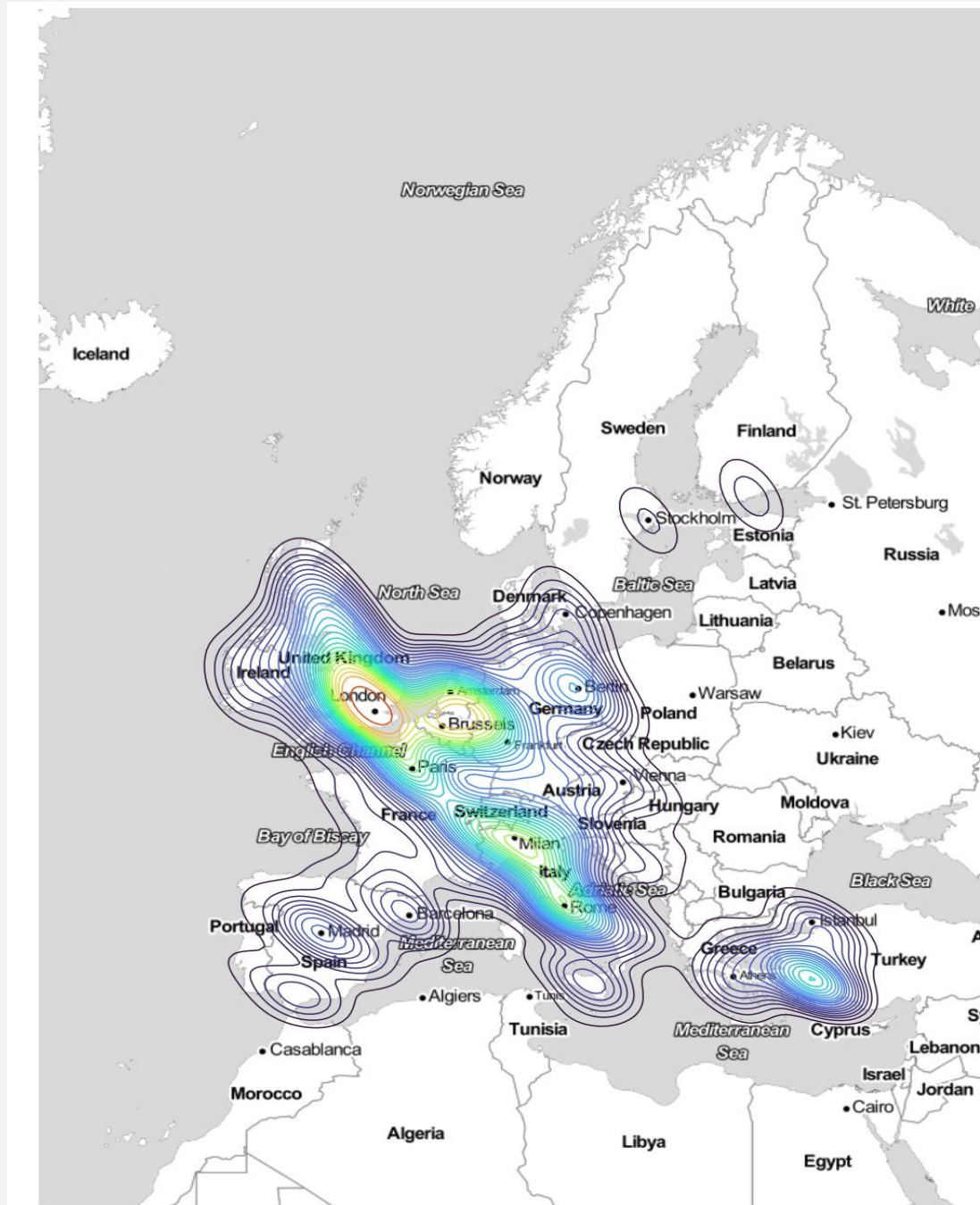- Facets of LBSN are disconnected, unlike normal Twitter data

Layout

- Introduction

- Objectives

- Background

- Objective 1

- **Objective 2**

- Objective 3

- Conclusion

23
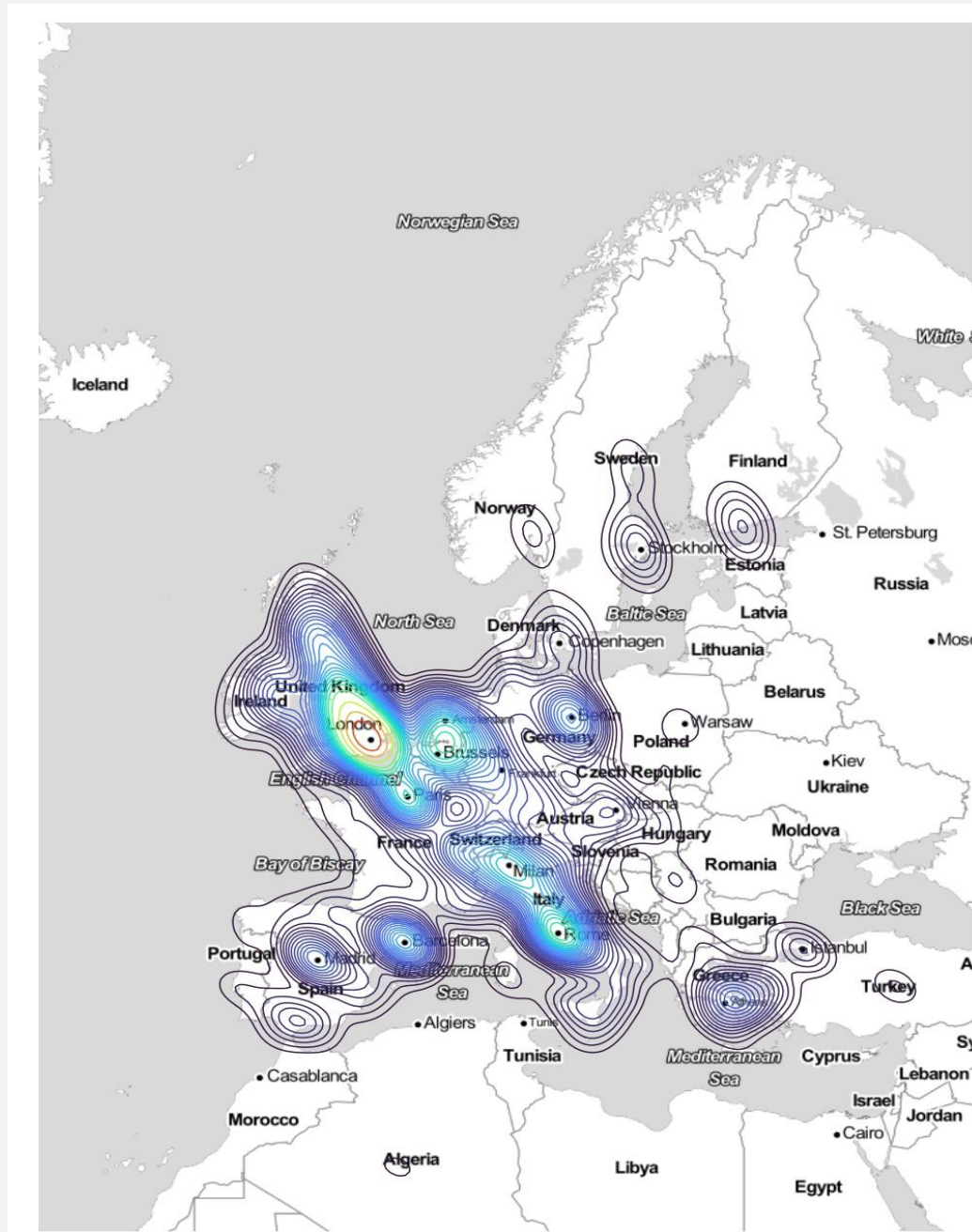
Objective 2: How does public reaction vary across time and space?
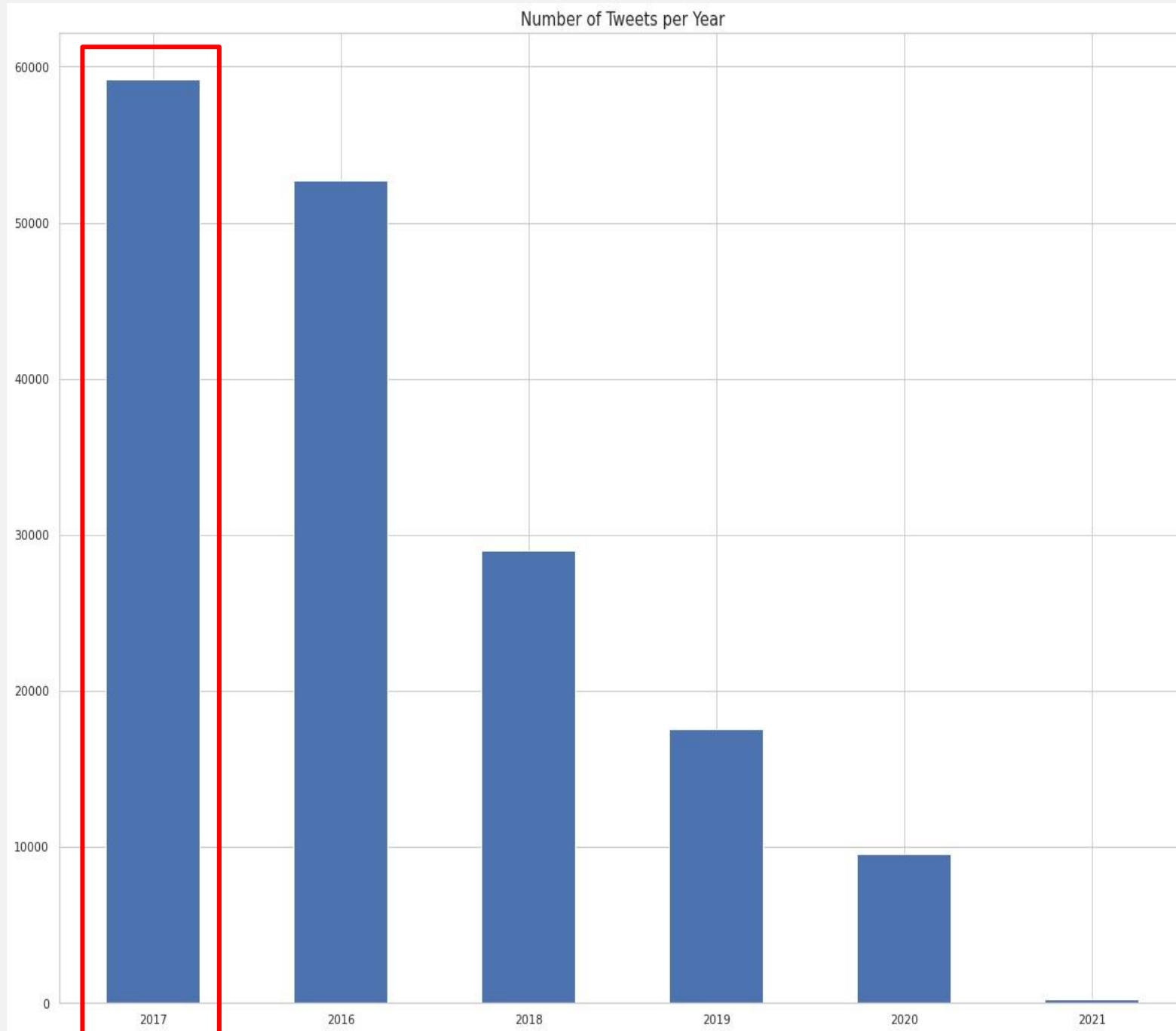
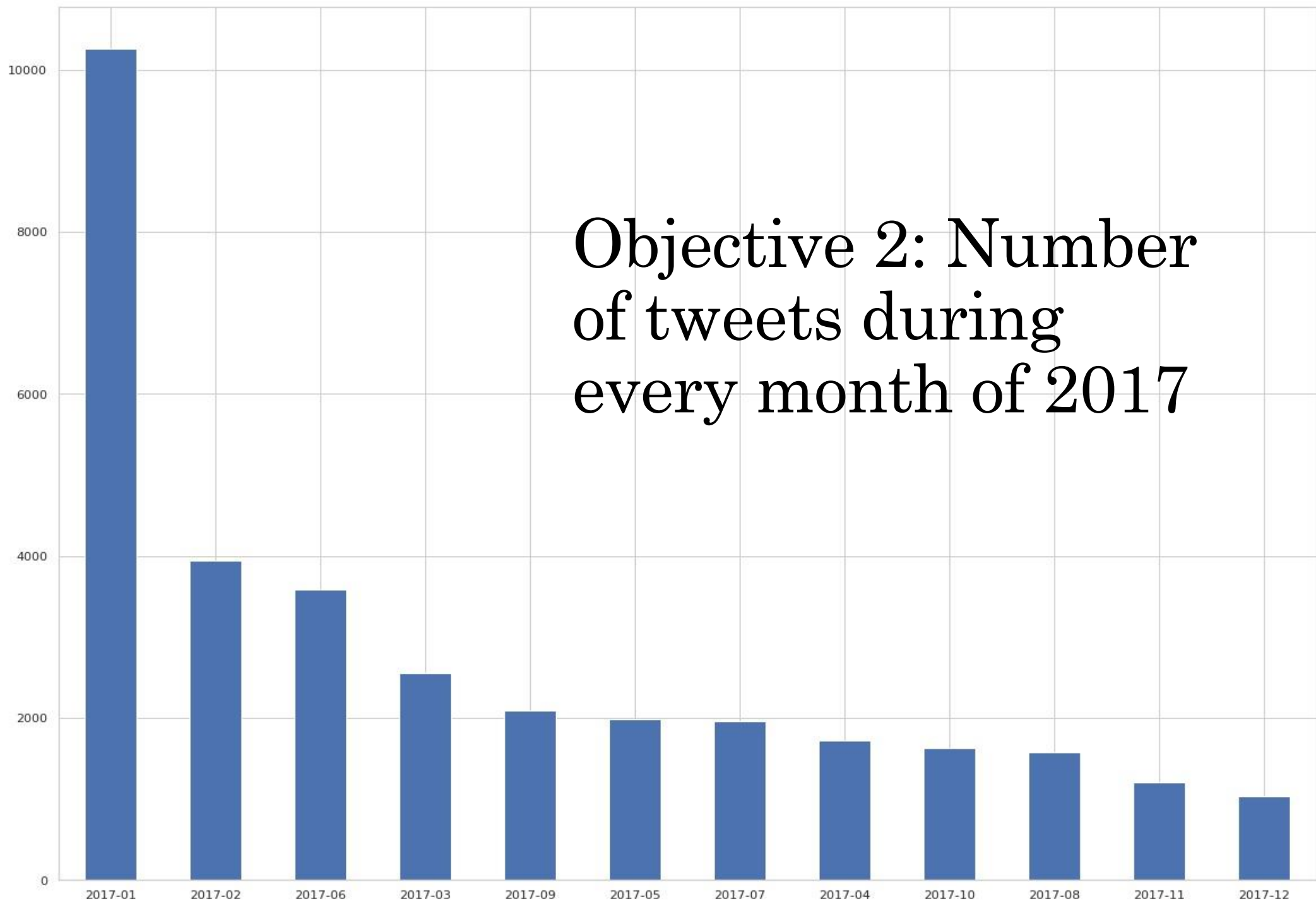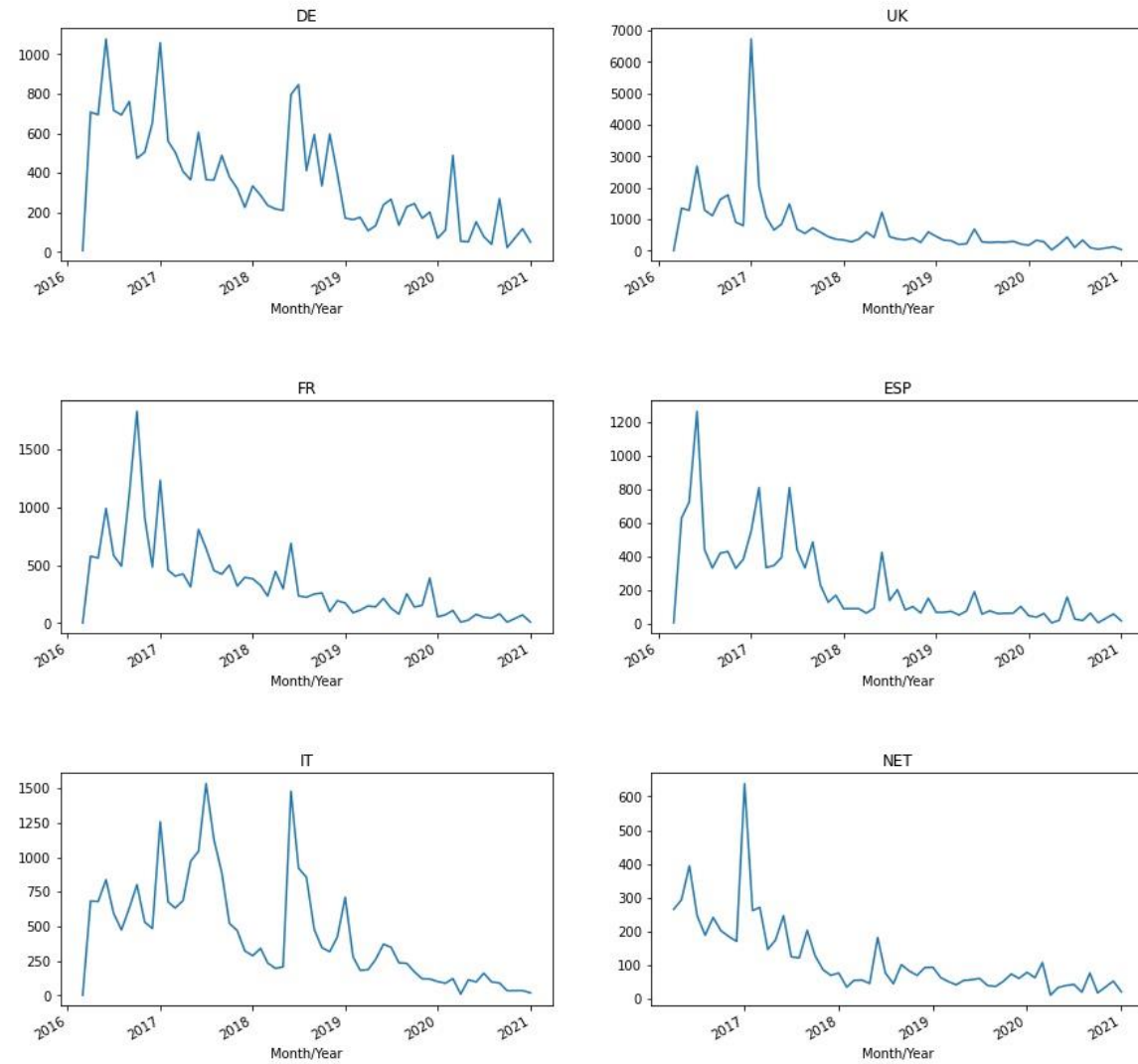Objective 2: How does public reaction vary across time and space?

Objective 2: How does public reaction vary across time and space in 2017?

Objective 2: Number of tweets during every month of 2017

Tweets per year per country

Objective 2: Temporal, Spatial and Topical

Hashtags during peaks in Germany

- Spatial : Country Specific

- Temporal : Months of 2017

- Topical : Hashtags used during the peaks

Final Visualization: Dashboard with Streamlit

- Python library to make simple dashboards

- Supports various kinds of visualization

- Allows for interactivity

- Supports high-level mapping libraries for interactive maps

Objective 2 : Final Thoughts

- Twitter data is not sufficient for a complete narrative visualization

- Strict definition and designing based on a user group

- User evaluation to gauge the effectiveness of such a visualization

Layout

- Introduction

- Objectives

- Objective 1

- Objective 2

- **Objective 3**

- Conclusion

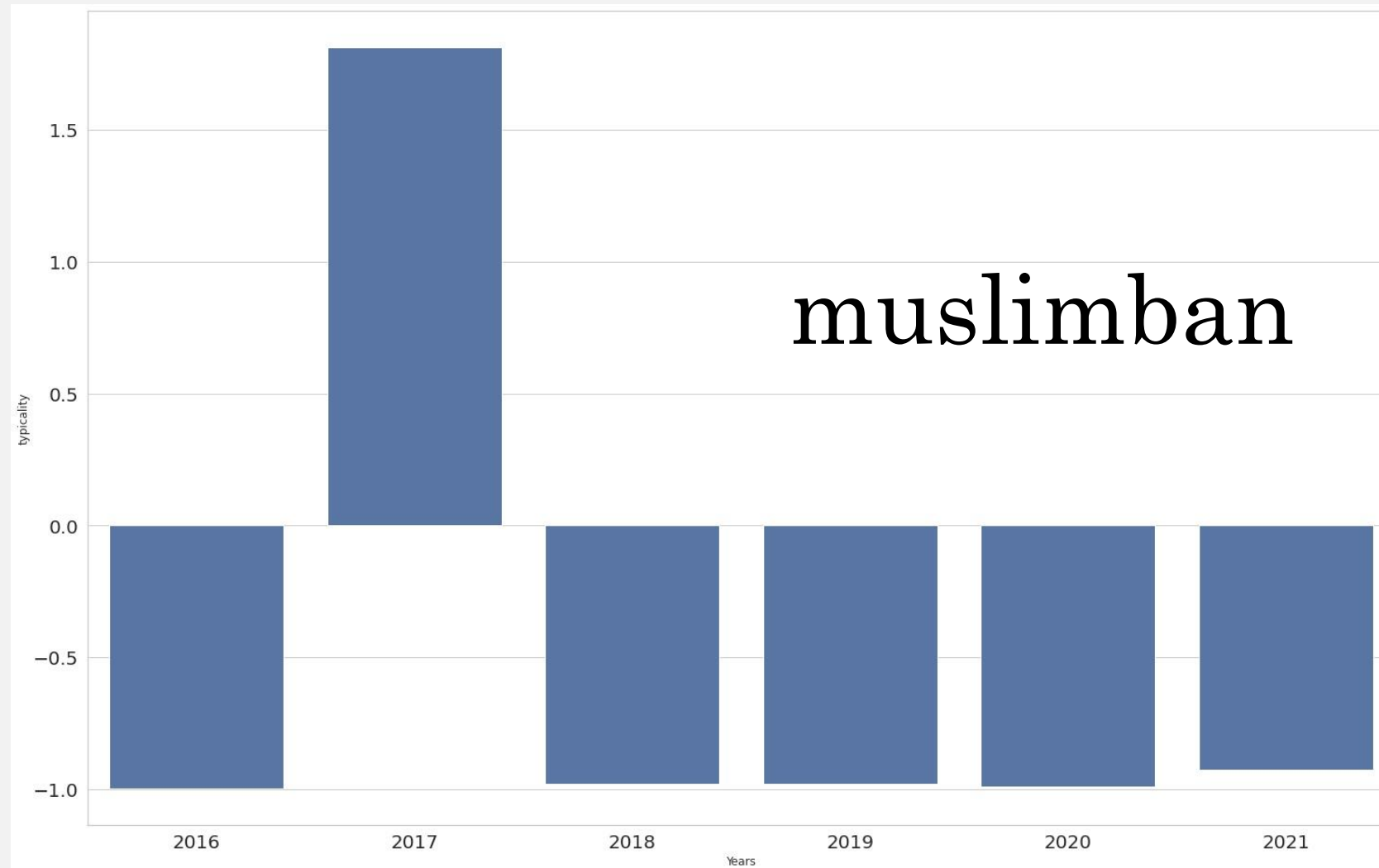# Objective 3 : Methodology

- Use hashtags to filter relevant tweets from database

- Second step of filtering to remove semantically non relevant tweets

- Final dataset

33

Objective 3 : Non-relevant hashtags/tweets

- ***asyl*** (German) picked up e***asyl***ike (English)

- ***moria*** picked up ***moria***rty

Objective 3 : Hashtags and Events

muslimban

Objective 3 : Co-occurring hashtags

Objective 3 : Hashtags and Events

refugeeswelcome(rw)
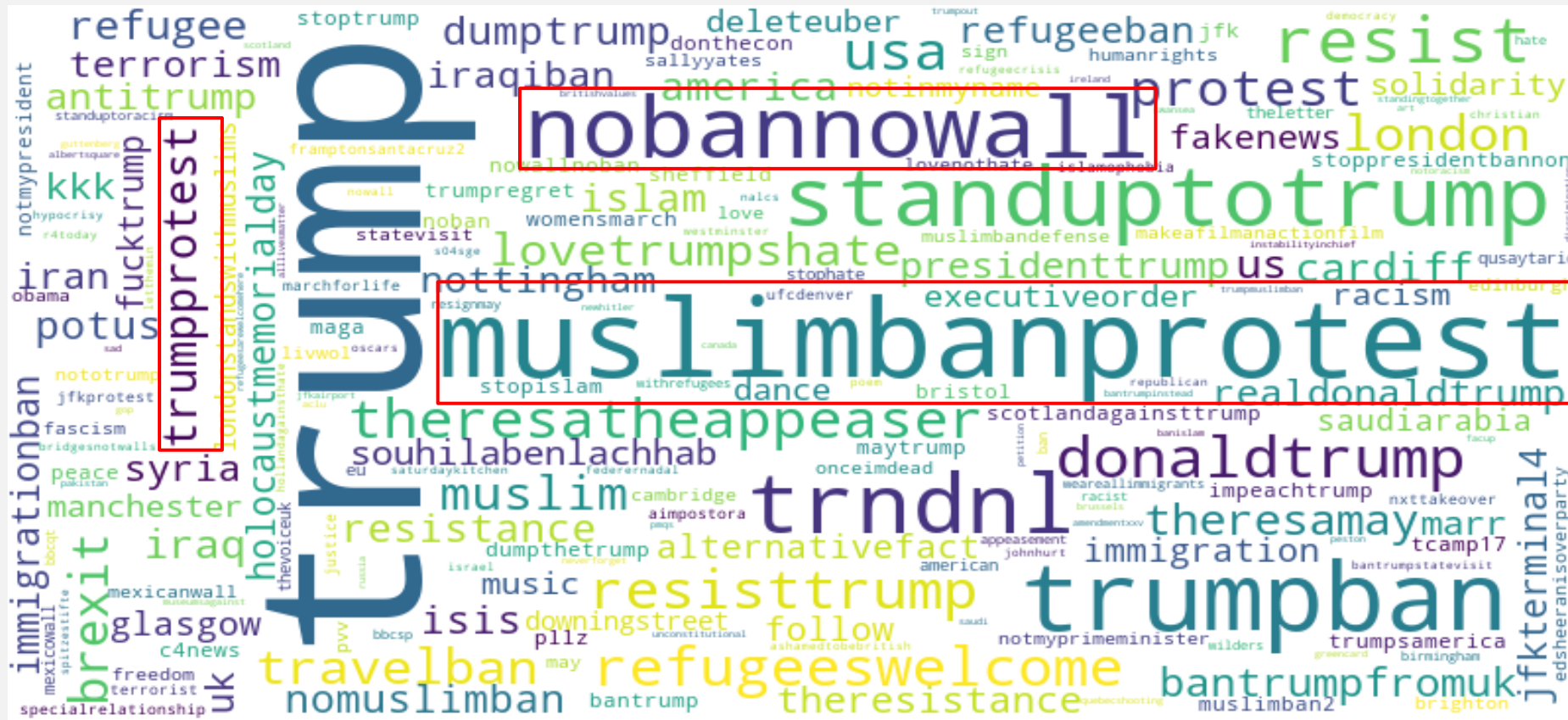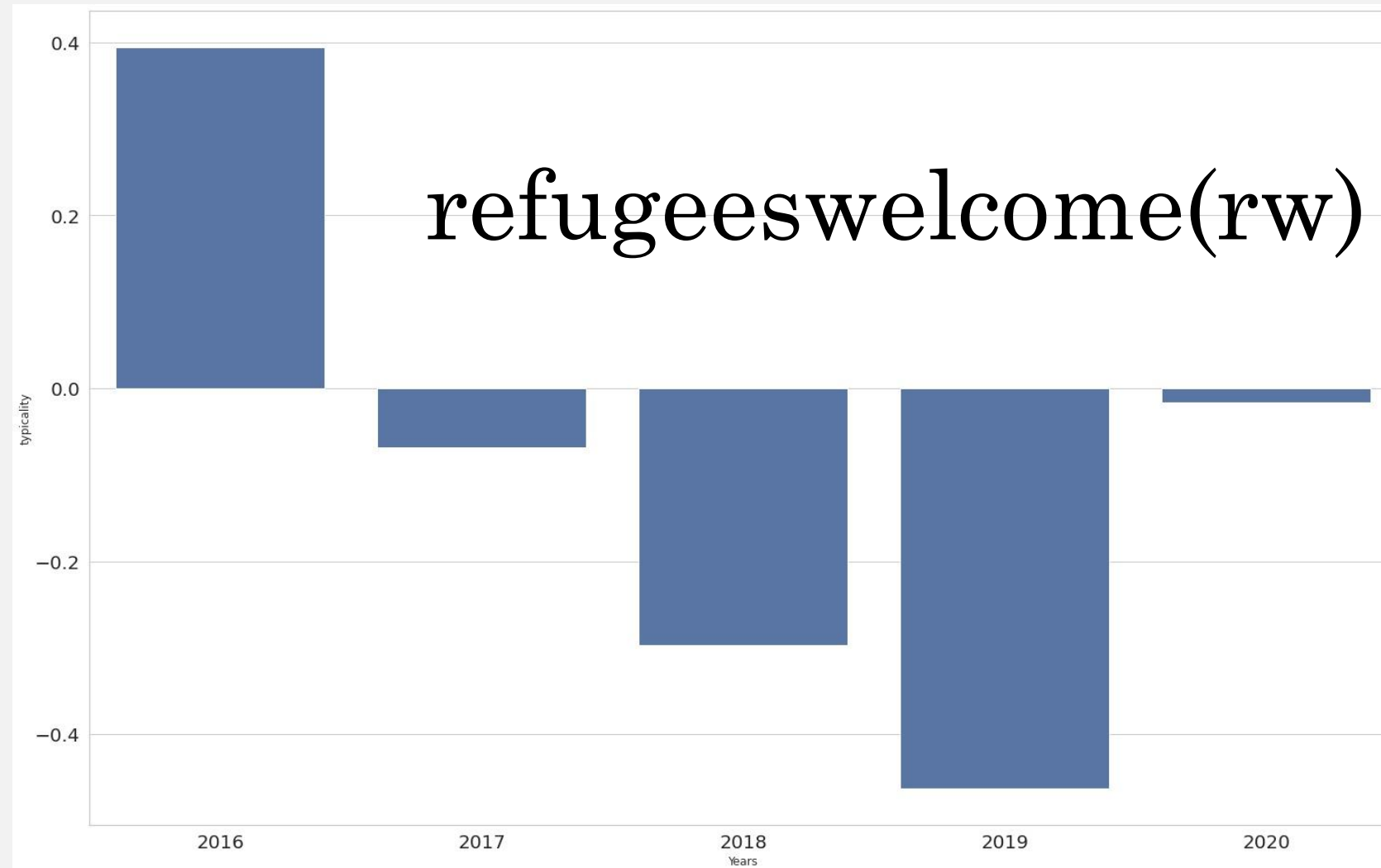
Co-occurrence of rw in entire dataset
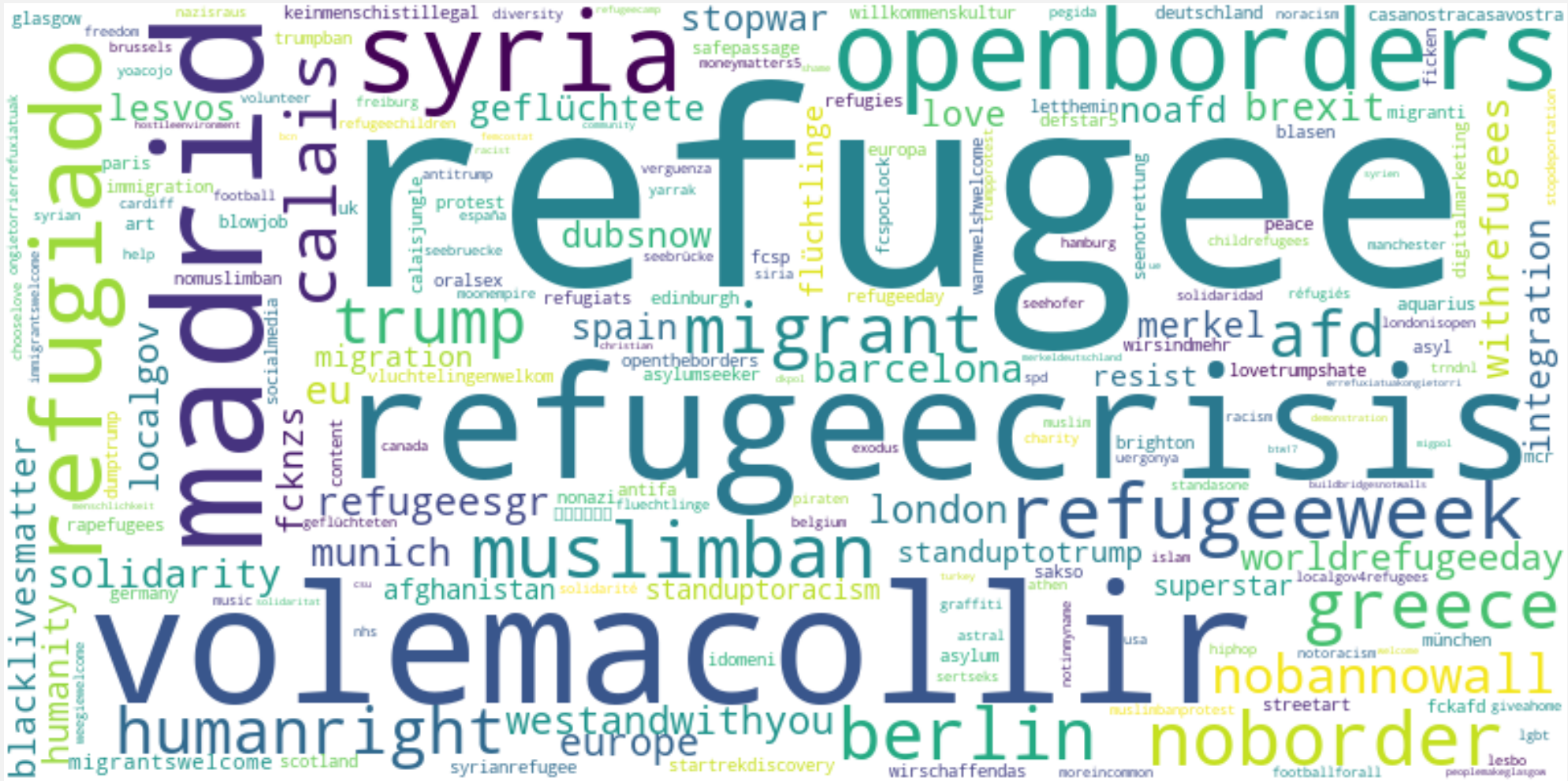
Co-occurrence of rw in 2016, Germany

Objective 3: Co-occurrence and opinion

rapefugees

Objective 3: Ambiguity

Objective 3: Final thoughts

- Using hashtags for opinion analysis is simple

- Very helpful for multi-lingual datasets

- Making interpretations require case by case discretion and caution

Conclusion: The refugee crisis on Twitter

- Difficult to make conclusive remarks because of the broad nature of the

  crisis

- Reactions tied to events in the political landscape of the country

- Needs further sources of data to place the tweets in context

References

Dix, Alan; Ellis, Geoffrey (1998): Starting simple. In Stefano Levialdi, Tiziana Catarci, Maria Francesca Costabile, Giuseppe Santucci, Laura Taranfino (Eds.): Proceedings of the working conference on Advanced visual interfaces - AVI '98. the working conference. L'Aquila, Italy, 24/05/1998 - 27/05/1998. New York, New York, USA: ACM Press, p. 124.

Dunkel, A., Andrienko, G., Andrienko, N., Burghardt, D., Hauthal, E., & Purves, R.   (2019). A conceptual framework for studying collective reactions to events in location-based social media. *International Journal of Geographical Information Science*, *33*(4), 780–804. https://doi.org/10.1080/13658816.2018.1546390

Heer, Jeffrey; van Ham, Frank; Carpendale, Sheelagh; Weaver, Chris; Isenberg, Petra (2008): Creation and Collaboration: Engaging New Audiences for Information Visualization. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell et al. (Eds.): Information Visualization, vol. 4950. Berlin, Heidelberg: Springer Berlin Heidelberg (Lecture Notes in Computer Science), pp. 92–133.

Kumar, Shamanth; Morstatter, Fred; Liu, Huan (2014): Visualizing Twitter Data. In Shamanth Kumar, Fred Morstatter, Huan Liu (Eds.): Twitter Data Analytics. New York, NY: Springer New York (SpringerBriefs in Computer Science), pp. 49–69

Flajolet, P., Fusy, É., Gandouet, O., & Meunier, F. (2007, June). Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. In *Discrete Mathematics and Theoretical Computer Science* (pp. 137-156). Discrete Mathematics and Theoretical Computer Science.

References

Dunkel, Alexander; Löchner, Marc; Burghardt, Dirk (2020): Privacy-Aware Visualization of Volunteered Geographic Information (VGI) to Analyze Spatial Activity: A Benchmark Implementation. In IJGI 9 (10), p. 607. DOI: 10.3390/ijgi9100607.

Hauthal, E., Dunkel, A., & Burghardt, D. (2021). Emojis as Contextual Indicants in Location-Based Social Media Posts. *ISPRS International Journal of Geo-Information*, *10*(6), 407. MDPI AG. Retrieved from http://dx.doi.org/10.3390/ijgi10060407

Hauthal, E., Burghardt, D., & Dunkel, A. (2019). Analyzing and Visualizing Emotional Reactions Expressed by Emojis in Location-Based Social Media. ISPRS International Journal of Geo-Information, 8(3), 113. https://doi.org/10.3390/ijgi8030113

Thank you for your attention