



Cartography M.Sc.

Master thesis

Emojis as Indicators of Spatial-Temporal-Thematic Developments in Geo-Social Media

Samantha Levi



2022

Emojis as Indicators of Spatial-Temporal-Thematic Developments in Geo-Social Media

Submitted for the academic degree of Master of Science (M.Sc.)
Conducted at the Institute of Cartography, Department of Geosciences
Technical University of Dresden

Author: Samantha Levi
Study course: Cartography M.Sc.
Supervisors: Dr.-Ing. Eva Hauthal (TUD), Sagnik Mukherjee M.Sc. (TUD)
Reviewer: Dr. Frank Ostermann (UT)

Chair of the Thesis
Assessment Board: Prof. Dipl.-Phys. Dr.-Ing. habil. Dirk Burghardt (TUD)

Date of submission: 09.09.2022

Statement of Authorship

Herewith I declare that I am the sole author of the submitted Master's thesis entitled:

"Emojis as Indicators of Spatial-Temporal-Thematic Developments in Geo-Social Media"

I have fully referenced the ideas and work of others, whether published or unpublished. Literal or analogous citations are clearly marked as such.

Dresden, 9 September 2022

Samantha Levi

Table of Contents

1	Introduction	3
1.1	Motivation and Problem Statement	
1.2	Research Objectives and Questions	
1.3	Innovations Intended	
1.4	Thesis Structure	
2	Background	7
2.1	Geo-Social Media	
2.2	Emojis	
2.3	Facets of Location-Based Social Media	
2.4	HyperLogLog	
2.5	Typicality	
2.6	Visualizations	
3	Methods and Data Analysis	13
3.1	Data Collection and Preprocessing	
3.1.1	Twitter Data	
3.1.2	Emoji Processing and Simplification	
3.1.3	Country Data	
3.2	Raw Data Analysis	
3.2.1	Assessment of Data Completeness	
3.2.2	Data Exploration	
3.2.3	Temporal Typicality of Popular Emojis	
3.2.4	Spatial Typicality	
3.3	HLL Data Analysis	
3.3.1	Calculation of User Days per Emoji	
3.3.2	Visualization of User Days	
3.3.3	Calculation of User Days per Country	
3.4	Emoji-Specific Analysis	
3.4.1	Topical Consistency	
3.4.2	Temporal Typicality	
3.4.3	Spatial Typicality	
3.4.4	Spatial-Temporal Typicality	
3.4.5	Emoji-Specific Interpretation	
4	Results and Discussion	53
4.1	Research Objective 1	
4.2	Research Objective 2	
4.3	Research Objective 3	
5	Conclusion	57
5.1	Future Work	

References	59
Appendices	63
Appendix A Methods and Data Analysis	63
Appendix B Raw Data	64
B.1 Temporal Typicality for Top 50 Emojis by Absolute Frequency	64
B.2 Temporal Typicality for Top 50 Emojis by User Days	66
Appendix C HLL Code	68
C.1 Cardinality Function	68
C.2 Union Function	68
Appendix D Emoji-Specific Analysis	70
D.1 Summary of Analyzed Emojis	70
D.2 Ballot Box with Ballot Spatial-Temporal Matrix	71
D.3 Breastfeeding Spatial-Temporal Matrix	72
D.4 Clapping Hands Spatial-Temporal Matrix	73
D.5 Golf Spatial-Temporal Matrix	74
D.6 Hospital Spatial-Temporal Matrix	75
D.7 Mask Spatial-Temporal Matrix	76
D.8 Microbe Spatial-Temporal Matrix	77
D.9 Rainbow Spatial-Temporal Matrix	78
D.10 Raised Fist Spatial-Temporal Matrix	79
D.11 Woman Health Worker Spatial-Temporal Matrix	80
Appendix E Supplementary Materials	81

List of Figures

1 The mathematical definition of typicality	10
2 The effect of various aggregation levels on post locations.	16
3 Using the emoji Python library to remove skin tones modifiers from emojis. .	18
4 The spatial extent of the Twitter dataset. Country boundaries are used as a basemap in future visualizations	20
5 The number of posts collected per week in the total dataset. Some weeks not included due to lack of data	21
6 The top 50 most frequently used emojis across the entire dataset. Emoji size is proportionate to the frequency of use of each emoji	22
7 A word-cloud consisting of the top 150 hashtags used across the entire dataset. Text size is proportionate to the frequency of use of each hashtag.	23

8	A comparison of the spatial typicality of the folded hands emoji (🙏) visualized on a 50 by 50 kilometer grid (left) and a 100 by 100 kilometer grid (right)	28
9	A comparison of the spatial typicality of the microbe emoji (🦠) in cropped (left) and uncropped (right) formats	29
10	The grid generated for spatial analysis and visualization. Each cell represents an area of 10,000 square kilometers	30
11	The spatial typicality of the party popper (🎉) and hot beverage (☕) emojis. Both of these emojis demonstrate ambiguous spatial trends that do not give concrete insights into the use of either emoji.	31
12	The spatial typicality of the dog face (🐶) and soccer ball (⚽) emojis. Both of these emojis demonstrate distinctive, country-based spatial trends.	31
13	The spatial typicality of the water wave (🌊) and snowflake (❄️) emojis. Both of these emojis demonstrate distinctive, environment-based spatial trends	32
14	Unions were performed on emojis with multiple possible skin tones. Then, the number of user days per generic emoji was calculated	35
15	The number of user days (days where distinct users posted at least once) per 100 by 100 kilometer grid cell	37
16	The calculation of user days within the United Kingdom	38
17	The top 20 co-occurring hashtags associated with the clapping hands (👏) emoji	41
18	The number of tweets collected per month in the raw Twitter dataset. Some unavoidable gaps in data collection create inconsistencies from month to month. No data was collected for the month of November.	43
19	The absolute frequency of the face with medical mask emoji (😷). The significant drop after March is misleading due to a gap in the dataset during the month of April	44
20	Typicality of the face with medical mask emoji (😷) calculated using monthly subsets	44
21	Typicality of the beer mug emoji (🍺) calculated using 100 by 100 kilometer grid cells as spatial subsets	45
22	Typicality of the raised fist emoji (✊) over time	47
23	Typicality map of the raised fist emoji (✊)	48
24	Typicality map of the ballot box with ballot emoji (🗳)	49
25	Typicality over time of the breastfeeding emoji (🤱)	49
26	Typicality over time of the face with medical mask emoji (😷)	50
27	Typicality over time of the rainbow emoji (🌈)	50
28	Typicality map of the clapping hands emoji (👏)	52

List of Tables

1	The most frequently used emojis for each monthly subset of the data, excluding November due to lack of data. Emojis listed in descending order by occurrence number	23
2	The most frequently used emojis in each of the ten countries with the highest number of user days. Emojis listed in descending order by occurrence number	24
3	Differences in user day calculations due to the aggregation level at the time of data collection	39
4	An example of a co-occurring hashtag table with topically consistent hashtags highlighted in gray. This process was repeated for each of the 35 emojis that were analyzed for topical consistency	42

Acknowledgements

I'm incredibly fortunate to have many people to thank for their support during the massive undertaking that became this thesis 🙏. Firstly, my thanks goes to my first and second supervisors, Dr.-Ing. Eva Hauthal and Sagnik Mukherjee, for their moral and technical support and for lending me their expertise throughout my research process. I would also like to thank my thesis assessment board including Dr. Frank Ostermann and Prof. Dipl.-Phys. Dr.-Ing. habil. Dirk Burghardt for their valuable advice and feedback during my research proposal and midterm presentations. I also extend my thanks to Dr.-Ing. Alexander Dunkel, who lent his expertise and technical guidance to this project.

Pursuing a Master's thesis ten thousand kilometers away from home has presented many challenges as well as wonderful opportunities. I would be remiss not to acknowledge Juliane Cron and her continuous support navigating the endless pile of paperwork that results from living abroad 🤝. I would also like to thank all of the university coordinators who made transitions between universities over the course of the program as seamless as possible.

Staying motivated while studying in the midst of an ever-changing pandemic situation would not have been possible without the incredible support of the Cartography family ❤️. We have supported each other through thick and thin and I will forever be grateful for the true friendships I have made in the past two years 😊.

Another shout-out also goes to my wonderful childhood friends Grace O'Malley and Dani Tuchman who never fail to make me laugh, no matter the situation.

Last but not least, I would like to thank my parents, Carol and Sami Levi, my siblings, Sara and Victor, my twin sister Carolyn, her daughter Luna 🐶, and both of my lovely grandmothers for their constant moral support. I can always feel the love even from an ocean away 🌊.

Abstract

Social media is ubiquitous in the modern world and its use is ever-increasing. Similarly, the use of emojis within social media posts continues to surge. Geo-social media produces massive amounts of data that can provide powerful insights into thoughts and reactions of individual users as well as large groups of people. Currently, artificial intelligence is often used to derive meaning from the text content of social media posts. However, this technique meets obstacles in the form of spelling mistakes, grammatical errors, slang, and sarcasm. This study seeks to use emojis, which are mostly language-independent, as an alternative medium to detect relevant topics within a Twitter dataset that is not thematically pre-filtered.

This research aims to 1) detect spatial-temporal change in emoji usage over time, 2) investigate whether significant changes in emoji usage over time and space correspond with significant events, and 3) assess the suitability of existing geovisualization techniques and adapt selected methods to generate static or interactive maps. The dataset used consists of approximately 4 million Twitter posts that were geotagged within Europe during the year 2020. Both the original data and data in the privacy-aware HyperLogLog data format are leveraged to explore the social, spatial, temporal, and topical facets of the data. Emojis were found to correlate with relevant topics including the COVID-19 pandemic, several political movements, and leisure activities. The spatial and temporal developments of these topics were approximated through the variations in use of corresponding emojis. Ultimately, map matrices depicting the spatial-temporal evolution of emoji use were created for emojis that were found to be topically consistent.

Keywords: geo-social media, location-based social media, emoji, spatial-temporal analysis, typicality

1 Introduction

1.1 Motivation and Problem Statement

Social media plays an increasingly large role in the lives of over half of the world's population, with an additional 227 million new users creating accounts in the last year alone (Strick, 2022). Social media can therefore be seen not only as a platform through which users express thoughts and ideas, but also as a powerful source of data generated by billions of users. The content of social media posts can provide valuable insights into individual and, when analyzed in large quantities, collective reactions to events, products, and people (Gabarron, Dorronzoro, Rivera-Romero, & Wynn, 2019; K.-S. Kim, Kojima, & Ogawa, 2016; Kruspe, Häberle, Kuhn, & Zhu, 2020; C. Li, Sun, & Datta, 2012). General mood and public perception can also be gauged by analyzing social media data at large scales and, in the case of geotagged geo-social media data, can reveal thematic trends across space (Hauthal, Dunkel, & Burghardt, 2021; Imran, Daudpota, Kastrati, & Batra, 2020; K.-S. Kim et al., 2016).

Currently, geo-social media data is analyzed to inform everything from assessing customer satisfaction to predicting results of elections (Ayvaz & Shiha, 2017; Hauthal, Burghardt, & Dunkel, 2019). Doing so can lead to meaningful conclusions about current events and topics that attract large numbers of users (Andersson & Öhman, 2017; C. Li et al., 2012).

Many studies use text-based analysis approaches including Natural Language Processing to derive semantic meaning and detect topics of discussion in social media data. However, due to the inherent complexities of language, this method is subject to errors arising from jokes, sarcasm, negations, and slang (Hauthal et al., 2021). The quality of text-based analysis is further degraded when applied to social media data such as Twitter posts because of the prevalence of spelling, grammatical, and punctuation errors (Wiesław, 2016). Text-based analyses are also often restricted to a single language and are highly susceptible to inconsistencies arising from cultural differences (Kejriwal, Wang, Li, & Wang, 2021).

The analysis of emojis for the identification of local topics circumvents many of the downfalls of purely text-based analysis, since it can be conducted independent of language (Hauthal et al., 2019; Kejriwal et al., 2021). For example, the semantic meaning of emojis are less likely to be influenced by negation and slang than results derived from text-based analysis. Emojis are widespread, pervasive, and present in an ever-increasing portion of social media posts and communications (Barbieri, Espinosa-Anke, & Saggion, 2016; Ljubešić & Fišer, 2016). They are a valuable resource that can play a vital role in interpreting meaning expressed in social media data (Guibon, Ochs, & Bellot, 2016).

Social media data can already be seen as a supplement to official data sources; in the same way, emojis can be seen as an additional enriching supplement (Hauthal et al., 2019). However, despite the prevalence of emojis in social media (Barbieri et al., 2016; Broni, 2022), research regarding their use and interpretation remains limited

(Ayvaz & Shiha, 2017; Guibon et al., 2016). In studies that have incorporated emojis into text-based analyses of social media data, many have found that the addition of emojis has significantly enriched their results (Chen, Yuan, You, & Luo, 2018; Guibon et al., 2016; Kejriwal et al., 2021). However, studies that only look at the absolute frequency of emoji use are limited in their ability to differentiate regions from one another (Barbieri et al., 2016; Kejriwal et al., 2021). Typicality is a normalized statistical measure that has been proposed to solve this issue with promising results (Hauthal et al., 2021), however more research needs to be conducted in order to confirm the efficacy and reliability of this measure.

This study therefore seeks to determine the ability of emoji-based analysis to identify both relevant topics and their spatial-temporal evolution within a Twitter dataset. K.-S. Kim et al. (2016) proposed a methodology for identifying local topics in geo-social media using latent spatial-temporal relationships. Similarly, this thesis uses emojis contextualized with hashtags to identify topics within the dataset. As in Kruspe et al. (2020), the dataset will be unfiltered by topic in order to explore the full thematic diversity of the dataset. Typicality calculations will be implemented and compared with other standard statistical measures to derive meaningful results.

1.2 Research Objectives and Questions

The overarching research objective for this study is to investigate the extent to which emojis can be used to identify relevant topics as well as their spatial-temporal evolution in a geo-social media dataset that is originally not pre-filtered thematically.

Although emojis have not yet been widely researched within the realm of cartography, studies from other fields such as linguistics and computer science can provide meaningful insights into the way geo-social media posts, and more specifically the emojis within them, are used to express thoughts and ideas across time and space. Based on a thorough review of such existing literature, the following Research Objectives (RO) were selected for exploration:

- RO1:** Develop a means for detecting change in emoji usage over time.
- RO2:** Determine whether significant changes in emoji usage over time and space correlate with significant topics.
- RO3:** Visualize the results in a meaningful and comprehensive way.

Whereas other studies select specific topics and filter their datasets to contain only thematically relevant tweets (Andersson & Öhman, 2017; Chandra & Krishna, 2021; Gabarron et al., 2019; Imran et al., 2020; Mukherjee, 2021), this analysis will purposefully use a dataset unfiltered by topic to detect changes in general emoji usage over time and space and across topics. This research does not follow the traditional workflow of a sentiment analysis and is not directly intended as such. Rather than revealing the emotions or opinions of large groups of people, this study seeks to identify relevant topics discussed within the dataset.

The three main research objectives in this project each have corresponding Research Questions (RQ) that should be answered in order to achieve the overarching goal of the project.

- RQ1:** Do significant changes in emoji use happen over time and space?
- RQ2:** Do spatial and temporal changes in emoji usage have thematic connections?
- RQ3:** What are the most appropriate visualization methods to represent these thematic, temporal and spatial changes?

The conducted analysis should describe changes in emoji usage over time and space if they exist. The suitability of emojis for answering the given research questions should also be assessed and the advantages, disadvantages, and special qualifications of the approach should be discussed.

1.3 Innovations Intended

Emojis have been widely studied in disciplines including computer science and linguistics, but they are still under-researched within the realm of cartography. This study contributes to the broader understanding of emojis within the applications of cartography. Most studies that currently exist on the use of emojis in geo-social media aim to analyze emoji use either over time or space, but seldom both. In this study, a three-fold analysis will be conducted to identify trends in emoji usage over time, space, and space-time.

Furthermore, this study uses an exploratory approach to analyze trends across topics, rather than filtering the dataset to restrict analysis to a specific topic as is often done in existing literature. This approach differs significantly from the vast majority of studies analyzing spatial-temporal changes in emoji use; most of these existing studies analyze emoji usage only as it pertains to a particular subject and use keywords and other filters to narrow the thematic diversity of the dataset.

Essentially, the research performed in this thesis is notable for the dataset, which is not pre-filtered based on topic, and for the threefold analysis of this dataset at temporal, spatial, and spatial-temporal dimensions. The visualization of trends in emoji usage over both time and space is seldom found in existing literature. The final visualizations of this research are therefore also novel contributions.

1.4 Thesis Structure

The thesis will begin with a thorough review of existing literature pertaining to several main concepts present in the study. Section 3 will provide a brief introduction to the datasets used. Then the three main sections of the workflow will be presented in roughly chronological order and presented alongside specific results, including emoji-specific trends and analysis. These specific results will be synthesized and contextualized with regard to the posed research questions in Section 4. This section will also include a

discussion of the relative benefits, limitations, and necessary considerations of the methodology. Finally, the thesis will conclude with a summary of major findings and recommendations for further investigation. A link to supplementary materials including Python code, results, visualizations, and a PDF version of this thesis can be found in Appendix E.

2 Background

2.1 Geo-Social Media

Social media platforms such as Twitter, Instagram, and Flickr are conduits for users to express their emotions and reactions to various ideas and current events (Dunkel et al., 2019; Kruspe et al., 2020). The subjective nature of social media makes it ideal for gaining insights into public opinions and sentiment, and indeed many studies exist that leverage social media data expressly for sentiment analysis in response to a given topic (Chandra & Krishna, 2021; Gabarron et al., 2019; Imran et al., 2020; Kruspe et al., 2020). In these contexts, social media allows for the collection of subjective and user-related information on a scale that would be impossible to replicate using traditional survey-based data collection methods (Hauthal et al., 2021).

Social media data has high volume, veracity, and velocity (McKittrick, Schuurman, & Crooks, 2022) which can act as both an advantage and disadvantage during analysis. One advantage is that the sheer volume of data allows for trends to be analyzed across large groups of people. Indeed, Dunkel et al. (2019) proposed a methodology for the characterization and comparison of collective reactions, which are defined as a series of posts referencing the same event or idea. When location information is available for individual posts, it is also possible to gain a spatial understanding of where reactions to a specific event occur and if they differ from place to place. Goodchild (2007) famously described how individuals act as sensors when they contribute volunteered geographic information (VGI) to public platforms like OpenStreetMaps. Similarly, social media data can be seen as a form of volunteered information when posted to a public interface like Flickr or Twitter.

While the majority of social media data is aspatial and contains no geographic information, some geo-social media posts contain geographic information when users decide to geotag their posts. Geotagging refers to the recording and subsequent publication of the exact coordinates of a user at the time the post is published. The coordinate information is stored as an attribute of the social media post within the post metadata. On Twitter, precise geolocation is turned off by default -only users who actively alter their settings and change this function are able to post geotagged posts. This requirement means that only about 0.85% of all tweets are geotagged and that this subset is not necessarily representative of the wider Twitter population or of the global population (Malik, Lamba, Nakos, & Pfeffer, 2021; Sloan & Morgan, 2015).

Although only a very small percentage of posts are geotagged, with approximately 200 billion tweets published per year, this fraction still encompasses approximately 1.7 billion posts per year. These 1.7 billion posts can be considered as passive crowdsourced geographical information (See et al., 2016). Passive crowdsourcing involves the opportunistic use of data which is generated by non-professionals and published on sites other than formal citizen science projects (Ghermandi & Sinclair, 2019). This includes information that is uploaded to web-based social media platforms like Twitter. Indeed,

geotagged Twitter data is an extremely popular source of volunteered geographic information (VGI); in a survey of 59 papers using VGI for disaster management, an overwhelming majority of data-centric studies were found to use Twitter as a data source.(Granell & Ostermann, 2016).

2.2 Emojis

Emoji is a Japanese word meaning picture character (Guibon et al., 2016). Originally developed in the 1990s by Shigetaka Kurita for Japanese mobile phone providers (Lin & Chen, 2018; Ljubešić & Fišer, 2016), emojis can represent not only faces, but also concepts and objects (Guibon et al., 2016). They are widely used in internet communication, and their popularity has even led to the laughing-crying emoji (😂) being named the Word of the Year by the Oxford English Dictionary in 2015 (Bai, Dan, Mu, & Yang, 2019). Their popularity is only increasing; both June and July 2022 were record-breaking months for emoji usage, with the highest instances of emojis-per-tweet on record. In June 2022, over 22% of all tweets contained at least one emoji (Broni, 2022).

Emojis are just one of many non-verbal cues used to aid computer-mediated communication, along with capitalization, exclamation points, and emoticons (Bai et al., 2019). It is essential here to note the difference between emojis, which are Unicode characters rendered differently depending on a device's operating system, and emoticons, which are ASCII character sequences that mostly resemble facial expressions either horizontally (Western) or vertically (Eastern) (Guibon et al., 2016; Wiesław, 2016). While both are considered paralinguistic cues (Prada et al., 2018), some emoticons have no emoji equivalent, and vice versa (Guibon et al., 2016).

Studies have indicated that emojis are not included in tweets arbitrarily. Feldman, Barach, Srinivasan, and Shaikh (2021) found that a compensatory relationship between emojis and lexical diversity exists; when posts not containing emoji were compared with posts containing emojis, emoji posts have less variation in vocabulary. M. Li, Chng, Chong, and See (2019) determined that emojis convey clear semantics that can be used to supplement and fill in the gaps from sentiment analyses conducted with Natural Language Processing. This suggests that emojis are not simply used as afterthoughts, but play a significant role in online communication.

2.3 Facets of Location-Based Social Media

The four facets of Location Based Social Networks (LBSN) as presented by Dunkel et al. (2019) provided a conceptual framework with which to understand reactions to events in geo-social media posts. The four facets are social, temporal, spatial, and topical. These facets refer to the user creating the post, the time at which the post was created, the geotagged location of the post, and the topic discussed in the post, respectively.

Individual facets can be analyzed on their own. For example, Sloan and Morgan (2015) specifically investigated the social facet by analyzing the factors that determine which

demographics are most likely to geotag their geo-social media posts. Malik et al. (2021) similarly identified distinct users in the contiguous United States to determine whether geotag users are representative of their surrounding population. Facets can also be analyzed in tandem; Barbieri et al. (2016) explored both the topical and spatial facets of Spanish emoji usage by investigating both the meanings of various emojis as well as their usage across Barcelona and Madrid. Y. Kim, ki Kim, Lee, woo Lee, and Andrade (2019) leveraged the social and spatial facets of a Flickr dataset to quantify the number of distinct users in various parks. In this way, even studies that do not directly reference the four facets framework of LBSN tend to implement this logical division of attributes for data analysis.

2.4 HyperLogLog

HyperLogLog (HLL) is an algorithm and data abstraction format that can be used to efficiently estimate the number of distinct elements in a large dataset. The process of determining unique elements is called cardinality estimation (Dunkel, Löchner, & Burghardt, 2020). HLL is often used for scenarios involving Big Data, when counting the number of distinct elements could require large amounts of memory. For example, in order to count the exact cardinality in a dataset of approximately 4 billion elements, gigabytes of memory would be required. In contrast, counting 4 billion distinct elements with HLL requires only 5 bits(Flajolet et al., 2016). This benefit comes with a sacrifice of 3 to 5 percent accuracy in the final cardinality calculations. Because of this significant advantage, HLL is currently used by Google, Facebook and Apple to make sense of increasing data collections (Dunkel et al., 2020). The mathematical mechanics of the algorithm are explained thoroughly in Flajolet et al. (2016).

When working with geo-social media data, it is essential to account for both user over-representation and non-human users including bots (McKittrick et al., 2022). HLL helps to address both of these obstacles by reducing the influence of overactive users in the dataset. By counting only distinct users, a user that posts 30 times per day, for example, is counted the same as a user that posts only once per day. This helps to avoid instances in which hyperactive users may be over-represented in the dataset due to the sheer number of tweets they generate over time.

While other forms of VGI traditionally require some intention by the author to contribute to a scientific platform (Goodchild, 2007), geo-social media data, even when voluntarily geotagged, is often posted without this intent and explicit consent for data collection and analysis is usually missing (Löchner, Dunkel, & Burghardt, 2019). Therefore, although geo-social media data is publicly available, personal information should at least be abstracted to obscure sensitive user information since the data is being used in a way not originally intended by the author (Dunkel et al., 2020). HLL therefore suits research purposes in which absolute anonymity is not a requirement. Indeed, Desfontaines, Lochbihler, and Basin (2019) investigated the privacy retention of cardinality estimators like HLL and found that total privacy preservation using HLL is incompatible with accurate aggregation. It is therefore inaccurate to refer to HLL as a privacy-preserving algorithm.

Rather, the HLL data format can be used in combination with other data abstraction strategies, such as cryptographic hashing and spatial data aggregation, to improve the protection of private user information in comparison with raw data.

2.5 Typicality

Absolute and relative frequency are popular tools in studies that work with social media data. However, although many existing studies make use of these standard statistical measures for data analysis, many also point out that these metrics are sufficient only for general trend analysis and are limited in their use with emojis. This is because, although the absolute or relative frequency of word occurrence may offer some meaningful results, emojis are vastly more limited in their diversity and therefore the absolute and relative frequency of their occurrence provides results of only limited significance (Hauthal et al., 2021).

For example, M. Li et al. (2019) used absolute frequency to determine the most popular emojis used on Twitter globally and by country. Although this study was able to find some differences in emoji usage by country, the use of absolute frequency was insufficient for identifying other influential factors for emoji usage, such as the semantic meaning of emojis or the thematic context of the posts containing the emojis. In an analysis of emoji usage on Twitter in both Barcelona and Madrid, Barbieri et al. (2016) demonstrated that very few differences between emoji usage in the two cities were able to be revealed with the use of absolute frequency. Both studies highlight the insufficiency of absolute frequency for analysis over space and turned to other statistical measures including nearest-neighbor calculations in order to contextualize emoji usage.

Typicality is an alternative statistical measure that determines how typical a given emoji is within a subset of a given dataset. This measure was first introduced by Hauthal et al. (2021) and is calculated as the normalized difference of relative frequencies and the result is a proportional number indicating the typicality of the emoji occurrence. Figure 1 shows the equation for typicality, where n_s represents the number of occurrences of a specific emoji in a subset, N_s represents the number of all emojis in the subset, n_t represents the number of a specific emoji in the total dataset, and N_t represents the number of all emojis in the total dataset.

$$t = \frac{(n_s/N_s) - (n_t/N_t)}{n_t/N_t}$$

Figure 1: The mathematical definition of typicality.

If a given emoji's typicality is positive for a subset, the emoji is considered typical and if the typicality is negative, the instance is considered atypical. The greater the absolute value of the result, the more significant the result. Of course, other statistical measures

exist with which relative differences and normalization can be calculated, such as the tf-idf or chi-squared tests, but these metrics are "comparatively complex and partly bound to preconditions" (Hauthal et al., 2021). Typicality as a metric therefore offers a middle ground between simplistic absolute and relative frequency values and more sophisticated but computationally intensive measures.

2.6 Visualizations

Although data collection and analysis are integral to the overall output of research, visualizations often have the largest impact on reader communication (McKittrick et al., 2022). Effective visualization techniques are therefore essential for properly communicating the results of any study. In addition to basic visualization outputs such as graphs and tables, cartographic representations in the form of maps serve as significant vehicles of information communication due to their "visual nature and high information density" (McKittrick et al., 2022). Existing geovisualization techniques for social media data include point mapping, raster surfaces, thematic mapping, and transmission diagrams (McKittrick et al., 2022). Depending on the platform used for presentation, such visualizations can be either static or dynamic. Determining the most appropriate visualization technique depends largely on the type of data to be shown and the results that should be highlighted.

In the realm of emoji research, emojis have been combined with both basic and cartographic visualization techniques to convey information more efficiently. Chen et al. (2018), for example, used emojis as point labels in charts to convey the semantic value of various facial expression emojis. This visualization highlights semantic groupings of similar emojis and the emotions they convey. Similarly, in a study on the spatial differences of emoji usage, for example, Kejriwal et al. (2021) implemented a form of point mapping using emojis as points indicating the most commonly used emoji within a given region. This visualization highlights the trends in commonly used emojis and how the most commonly used emojis differ by location.

While using emojis as points can increase the information density of a map visualization, it is impractical for instances with large numbers of data points. Since many geo-social media datasets consist of millions or even billions or trillions of data points, other methods have been adapted to reduce both computing time and visual clutter. Choropleth maps in which pre-defined polygons are shaded depending on the quantity or characteristics of local geo-social media data present are one example of this. Koylu (2019), for example, implemented choropleth maps to visualize tweets containing keywords across the contiguous United States, allowing trends in the data to be visually assessed without the need for plotting individual locations of tweets.

In addition to thematic mapping using choropleth maps, spatial data aggregation and the reduction of spatial granularity can de-sensitize geo-social media data visualizations by further obscuring precise user locations. In some cases, the spatial data granularity may be altered between the data collection or visualization phases according to the needs of

the project. For example, data may be collected at a higher precision in order to improve the accuracy of the analysis but then generalized to a lower precision for visualization purposes in order to protect user privacy. In cases where the HyperLogLog algorithm is additionally used to de-sensitize a dataset, a new direction of visual analytics is enabled which is "specifically suited to exploration in combination with visualization techniques that focus on identifying patterns of data and contexts where definite answers are not a requirement" (Dunkel et al., 2020). Thematic mapping with the use of HLL data is therefore uniquely suited for visual exploratory data analysis.

3 Methods and Data Analysis

Three main analyses were ultimately conducted. The first two exploratory data analyses were conducted for the Twitter data in both its raw and HLL formats. These analyses were performed with the intention of gaining overarching insights about general emoji usage and to reveal initial trends in the data. Once results from these processes were gathered, a third, emoji-specific analysis was conducted to further investigate selected emojis that were found to be topically consistent. The general workflow is visualized in Appendix A.

In the interest of transparency and replicability, the precise steps of all data analysis and visualization are available as Jupyter Notebook files that can be found in this project's Github repository (see Appendix E).

3.1 Data Collection and Preprocessing

3.1.1 Twitter Data

Two main data formats were leveraged in this analysis: so-called "raw" Twitter data and Twitter data in the HLL format. All Twitter data used in this analysis was sourced by the Technical University of Dresden Institute of Cartography from the Twitter API as GeoJSON files. Twitter data can only be stored in a database as long as all attribute fields are preserved during data collection. In other words, location data cannot be stored or aggregated on a standalone basis (Twitter, Inc., 2020). Data was then converted from GeoJSON into CSV format and hosted on a remote server, where data was aggregated into tables based on the 4 facets of LBSN: social, spatial, topical, and temporal. For the purposes of this research, the terms "LBSN" and "geo-social media" can be used interchangeably. The term "raw" data is used only to compare the original data to the data in the HLL format. Raw data in this case is therefore modified from its original GeoJSON format by conversion into a CSV format and removal of unnecessary attribute fields. The raw data was used to produce HLL data via a custom schema whereby the data was also cryptographically hashed and geohashed.

HLL uses a non-cryptographic hash function called MurMurHash which stores information as character sequences called hashes. These hashes are structured into shards that contain pieces of information such as the number of distinct users in a given area (Löchner et al., 2019). While this structure allows for efficient cardinality calculations of unions and intersections of a dataset, such non-cryptographic hashing does not preserve privacy (Desfontaines et al., 2019). It is therefore practical to combine this measure with other data abstraction techniques like cryptographic hashing and geohashing.

Cryptographic hashing refers to the use of a one-way pseudonymization function to encrypt data values. This measure on its own is a somewhat superficial privacy measure, since the restricted length of fields like usernames means that attackers could potentially create rainbow tables of all possible hashed values of a specific user ID and then perform

a lookup (Yu & Weber, 2020). In combination with the non-cryptographic hashing used in HLL, however, cryptographic hashing is necessary for the prevention of so-called salt reconstruction attacks, whereby attackers learn all relevant parameters of the cardinality estimator (Desfontaines et al., 2019).

Geohashing is a means by which the accuracy of spatial coordinates is reduced to a fixed level of precision ranging from hundreds of kilometers to sub-meter accuracy (Valley, Usher, & Cook, 2017). Geohashing was performed on this project's HLL dataset using a geohash function that reduces the precision of coordinates based on a desired accuracy level in the form of an integer. Both datasets, sorted into tables according to facet, were ultimately stored in a remote server accessible via PgAdmin, a PostgreSQL management tool.

Both of these data types pose unique benefits for this analysis. The advantages of the raw data format are that the data is human-readable (due to the lack of cryptographic hashing) and that the data is intuitively structured with each row representing a single tweet. The raw data format was ideal for the simultaneous analysis of the temporal, topical, and spatial facets of the data since all of the necessary information was recorded for each post; each row contains spatial information in the form of coordinates, temporal information in the form of publication date and time, and topical information in the form of an emoji and hashtag combination. One major disadvantage of the raw data, however, is the lack of privacy-awareness. For example, although the number of distinct users in the dataset could be counted using the raw data, doing so would require the storage of usernames and would be quite computationally intensive. To preserve some level of privacy in the raw dataset, usernames were not sources at all from the remote server. Raw data was therefore not ideal for investigating the social facet of the given dataset.

HLL data, on the other hand, sacrifices some precision for the sake of privacy awareness. In combination with additional privacy measures like spatial data aggregation and cryptographic hashing, sensitive user information like usernames could be protected from all forms of attacks except for improbable intersection attacks which require the knowledge of the secret key used for hashing (Dunkel et al., 2020). These privacy benefits come as side-effects of the cardinality estimation capabilities of the algorithm.

The cardinality and union functions of HLL were also instrumental to reduce the reliance on absolute frequency as a metric during the raw data analysis; the calculation of user days per country and per emoji allowed the scope of the analysis to be narrowed down using a metric that is less influenced by hyperactive and non-human users than absolute frequency. However, data in the HLL format was less ideal for analysis of the temporal, spatial, and topical facets. Critically, the two components used to investigate the topical facet in this research -emojis and hashtags -were separated during the conversion from raw data to HLL data and could not be reunified due to the nature of the HLL format. Despite this limitation, the HLL format was ideal for both the desensitization of sensitive user information in the dataset and the analysis of the social facet of the data via user day estimation.

Because the aim of this research is to analyze data spanning a variety of topics, no keywords were used to topically filter the dataset. This crucial step deviates from the methodologies proposed in other analyses of emojis as indicators of thematic developments (Chandra & Krishna, 2021; Gabarron et al., 2019; Kruspe et al., 2020; Mukherjee, 2021) and is crucial to ensure that all topics of discussion can be analyzed over time and space. The data was only filtered to ensure that each post contained at least one emoji and one hashtag and was geotagged within Europe in the year 2020. The final raw dataset consisted of 4,020,046 rows, each representing a single post. The final HLL dataset consisted of 2,148,544 rows, each representing a unique combination of emoji and location. In the HLL dataset, a single row can represent multiple occurrences of the same emoji in the same location.

The SQL query shown in Code Listing 1 was used to select the raw data for downloading from the remote server. The data is sourced in the WGS 84 Web Mercator projection and is later re-projected into the Mollweide projection for visualization purposes.

```
SELECT
post_guid,
post_publish_date,
post_body,
hashtags,
emoji,
ST_X(ST_TRANSFORM(post_latlng,4326))ASLONG,
ST_Y(ST_TRANSFORM(post_latlng,4326))ASLAT
FROMtopical.post
ORDERBYorigin_idASC,
post_guidASC
```

Listing 1: Query used to source raw data from the remote server.

The HLL data was queried at multiple levels of spatial aggregation using a geohash function (Dunkel et al., 2020). Levels 3, 4, and 5 were queried and visually compared before a suitable spatial resolution was selected. The query shown in Code Listing 2 was used to collect the data at an aggregation level of 4. At this level, the spatial information is reduced in accuracy so that all aggregation points are approximately 19,545 meters away from each other. This measure further protects user information from intersection attacks (Dunkel et al., 2020).

```

SELECT
    _emoji_latlng.user_hll,
    _emoji_latlng.post_hll,
    _emoji_latlng.pud_hll,
    ST_Y(extensions.geohash_reduce(ST_SetSRID(ST_MakePoint(longitude,
        latitude),4326),4))As"latitude_4",
    ST_X(extensions.geohash_reduce(ST_SetSRID(ST_MakePoint(longitude,
        latitude),4326),4))As"longitude_4",
    _emoji_latlng.emoji,
    _emoji_latlng.latlng_geom
fromtopical._emoji_latlng;

```

Listing 2: Query to collect HLL data at an aggregation level of 4 using the geohash function.

Figure 2 below shows how changing the aggregation level of the data affects the spatial distribution of the data. At aggregation level 5, data points are accurate to approximately 3.8 kilometers. At aggregation level 4, they are accurate to approximately 20 kilometers, and at level 5, they are accurate to approximately 125 kilometers.

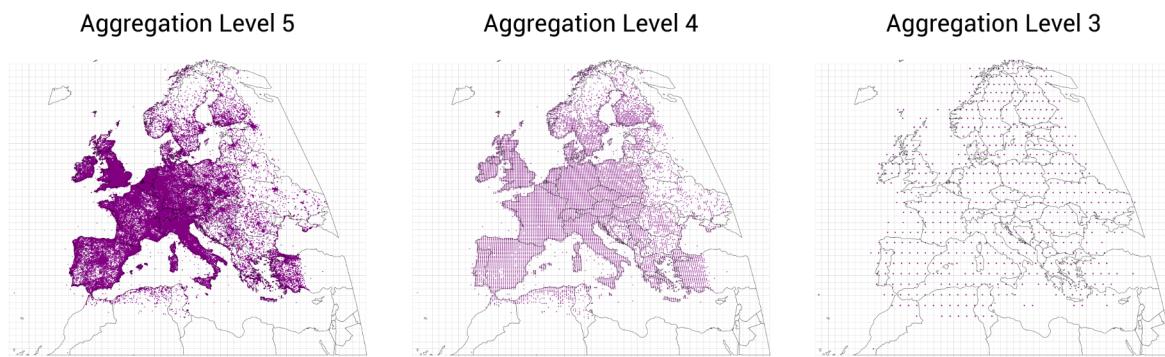


Figure 2: The effect of various aggregation levels on post locations.

In making a decision as to the level of spatial aggregation of the data, the Modifiable Areal Unit Problem (MAUP) was inevitable. The nature of the MAUP states that different levels of spatial data aggregation will lead to potentially different results during analysis (Openshaw, 1983). The MAUP is inherent to many geographical analyses but is particularly troublesome in studies involving social media data, since it is extremely difficult for researchers to be sure how to match the spatial unit of analysis to the scale of the phenomena being analyzed (de Andrade et al., 2021). In the case of this analysis, a spatial aggregation level of 4 was selected in order to provide a conservative level of spatial data desensitization while still allowing for visualization at an acceptable scale (see Section 3.3.2).

3.1.2 Emoji Processing and Simplification

While both datasets were originally filtered to only include posts that contain at least one emoji, some rows of the raw data attribute table had a blank emoji column after being read into the Jupyter Notebook. This is due to the representation of flags as combinations of Regional Indicator Symbol letters, which are alphabetic Unicode characters used to encode two-letter country codes. In the CSV-formatted raw data, these Regional Indicator Symbols are displayed as part of the post body attribute rather than as emoji attributes. Accessing these country codes in order to properly display flag emojis in the emoji attribute field would require storing the text content of each post in the raw and HLL dataset, which would significantly increase the amount of storage needed as well as raise potential privacy concerns. To avoid this, posts containing only flag emojis were removed from the dataset. This ensured that each row in the dataset contains at least one character in the emoji column. The step-by-step workflow of this process can be found in `RawDataCleaning_Final.ipynb` (see Supplementary Materials in Appendix E).

Another modification made to the raw dataset during data cleaning was the removal of skin tone modifiers from emojis that can be rendered with different skin tones. This was done so that multiple versions of the same emoji, for example the dark-skinned thumbs-up emoji (👍) and the light-skinned thumbs-up emoji (👎), could both be considered simply as the generic thumbs-up emoji (👉). For the purposes of this research, differences of skin color on the same emoji were considered to have negligible effects on the meaning of the emoji and were therefore removed to simplify calculations and analysis. For the raw data, skin tone modifiers were removed using the methodology shown in Figure 3. The resulting generic emojis were then stored in a separate column of the dataset and used in subsequent analysis.

```

# Let's convert the emojis to text to find the name
# of each skin tone
sample = emoji.demojize("👍👍👍👍👍👍")
sample

':thumbs_up_light_skin_tone: :thumbs_up_medium-light_skin_tone:
:thumbs_up_medium_skin_tone: :thumbs_up_medium-dark_skin_tone:
:thumbs_up_dark_skin_tone: :thumbs_up:'

# Let's remove the skin tone modifiers
sample = sample.replace("_light_skin_tone","");
sample = sample.replace("_medium-light_skin_tone","");
sample = sample.replace("_medium_skin_tone","");
sample = sample.replace("_medium-dark_skin_tone","");
sample = sample.replace("_dark_skin_tone","");
# and now convert the text back into emojis
sample = emoji.emojize(sample, language='alias')
# return the result
sample

'👍👍👍👍👍👍'

```

Figure 3: Using the emoji Python library to remove skin tones modifiers from emojis.

For the HLL data, a different method had to be implemented in order to eliminate the effect of skin tone modifiers on the analysis. In contrast to the methodology used with the raw Twitter data, there is no way to know whether two rows refer to two different posts or to one post containing two different emojis. For example, if one row describes a post containing at a location and another row describes a post containing at the same location, there is no way to know if these two rows refer to one or two separate tweets. While each of these two rows could represent separate posts, they could also represent two emojis used in the same post. Treating each row as its own post could result in an overestimation of the number of uses of the emoji. The methods used to mitigate the effect of skin tone modifiers for the HLL data are described in Section 3.3.

3.1.3 Country Data

In addition to the Twitter datasets, this research also required the use of a country boundary shapefile for both analysis and visualization purposes. Country boundary data was sourced from Natural Earth at a spatial resolution of 1:50 meters. In order to reduce the size of this file, many of the unnecessary attribute columns were removed. In the end, the only attributed that were preserved were the feature IDs, names of sovereign

nations, the formal name of each nation, the colloquial name in English, the shape area, shape length, and general geometry fields.

The SQL query shown in Code Listing 3 was then executed to find the bounding box for the given dataset - in other words, the maximum and minimum x and y coordinates of the dataset. For the purposes of this research, the study area was defined as the spatial extent of the available data and not according to any administrative boundaries. The study area will therefore generally be referred to as Europe although the area of analysis also includes small amounts of data from Middle Eastern, Asian, and North African countries.

```
SELECT
    MIN(ST_X(ST_TRANSFORM(post_latlng,4326)))asX1
    ,MIN(ST_Y(ST_TRANSFORM(post_latlng,4326)))asY1
    ,MAX(ST_X(ST_TRANSFORM(post_latlng,4326)))asX2
    ,MAX(ST_Y(ST_TRANSFORM(post_latlng,4326)))asY2
    from topical.post
```

Listing 3: Query used to find the spatial extent of the dataset.

The resulting bounding box is defined by the coordinates (-18.729512, 28.017169), (39.73858, 28.017169), (39.73858, 71.16987838), and (-18.729512, 71.16987838). In ArcGIS Pro, this bounding box was constructed and used to clip the global dataset of country boundaries to the spatial extent of the Twitter data (see Figure 4). The Mollweide projection was selected for visualization purposes throughout this thesis to avoid major size distortions in upper latitudes. The resulting clipped data was exported as a shapefile named Europe_Clipped_BBox.shp and was read into each of the relevant Jupyter Notebooks.

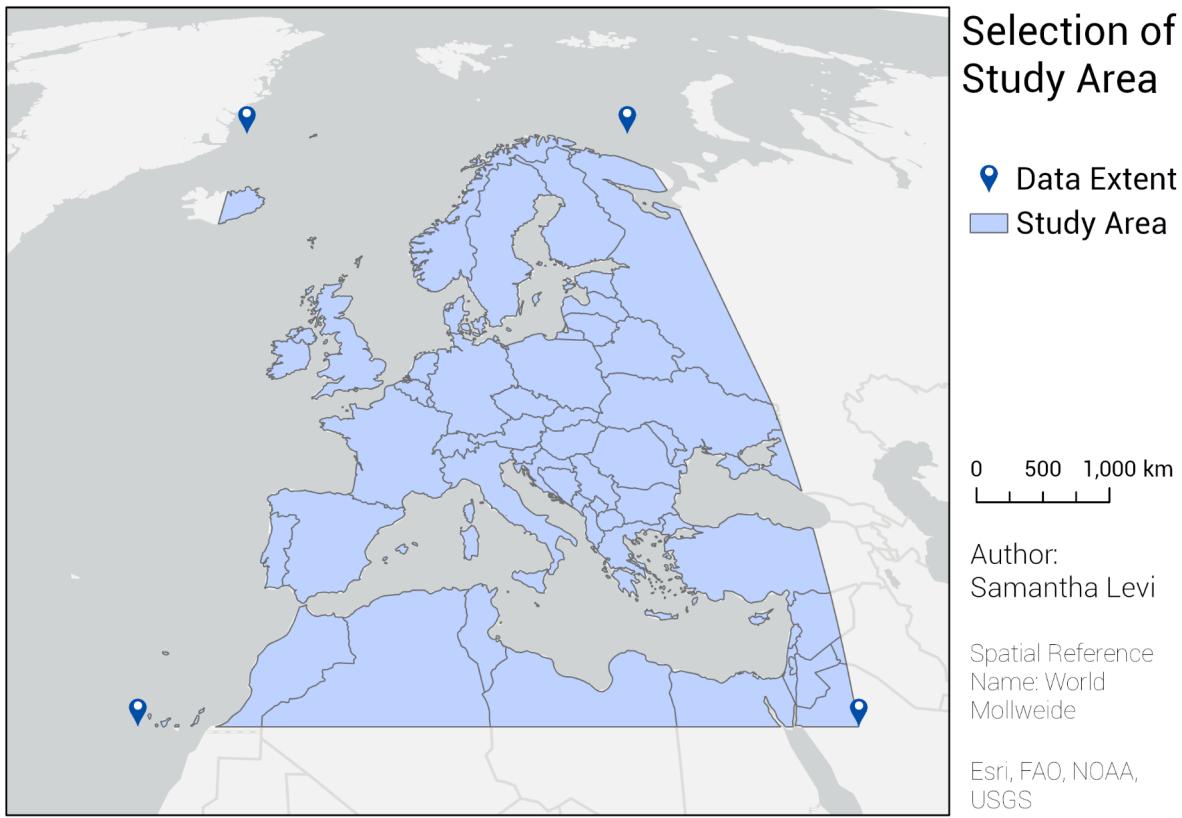


Figure 4: The spatial extent of the Twitter dataset. Country boundaries are used as a basemap in future visualizations.

3.2 Raw Data Analysis

The exact workflow and corresponding code for the following section are published in the Supplementary Materials in `Raw_DataAnalysis.ipynb` and `SpatialTypicality_Grid.ipynb` (see Appendix E). A condensed version of the iterative design process can also be found in the Supplementary Materials under `Spatial_Grid_Visualization_Experimentation.ipynb`.

3.2.1 Assessment of Data Completeness

After the raw data was prepared, the resulting GeoJSON file was read into a new Jupyter Notebook for exploratory analysis. Before commencing data analysis, it was essential to first investigate the quality of the dataset. According to the International Organization for Standardization, the components of data quality are completeness, logical consistency, spatial accuracy, thematic accuracy, temporal quality, and data usability (International Organization for Standardization, 2013). Of these components, logical consistency, spatial accuracy, and thematic accuracy were already addressed during the data cleaning and preparation stage. In order to determine the usability of the data, temporal quality was also considered.

Since the date and time of tweet publication were sourced directly from the Twitter API, they were assumed to be both accurate and precise. The temporal quality check

in this case was therefore mostly conducted in order to identify any temporal gaps in the data. To do this, additional columns were added to the dataset that aggregated the data into weekly, biweekly, and monthly intervals. Once the number of posts per month was rendered visually as a bar chart, it became obvious that certain gaps existed in the dataset, particularly in April and October (see Figure 18). In order to further investigate these gaps, the number of posts per week was also visualized (see Figure 5). Significant gaps were found for the second, third, and fourth weeks of April, the last two weeks of October, and the entire month of November. Unfortunately there was not sufficient time or resources to fill in these gaps since they had occurred at the time of data collection and additional data could not be retroactively sourced from the Twitter API. Given these unavoidable limitations in temporal quality, the data quality was still assessed as sufficient for the purposes of this research.

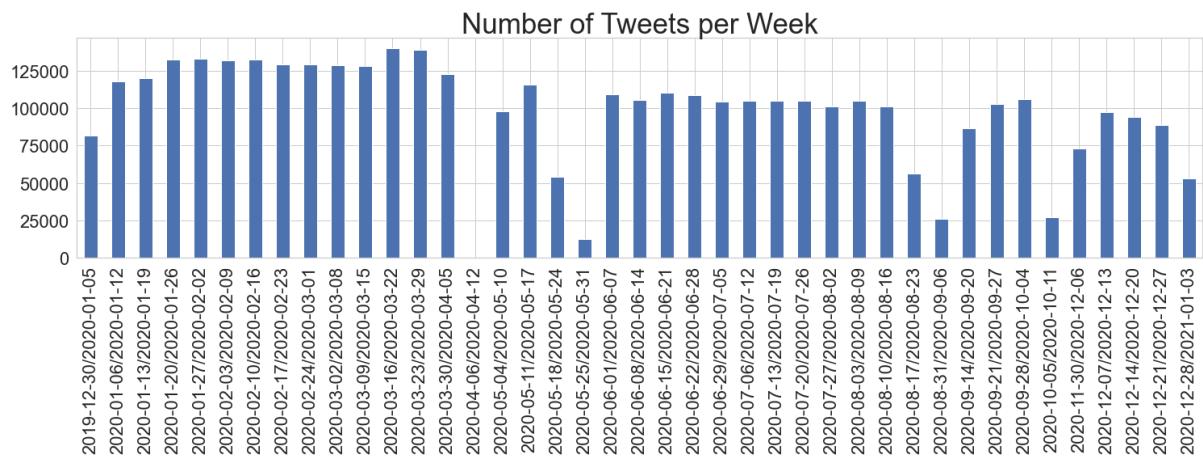


Figure 5: The number of posts collected per week in the total dataset. Some weeks not included due to lack of data.

3.2.2 Data Exploration

Once the degree of data quality was established, some visualizations were generated to gain an overall understanding of frequently used emojis in the given dataset. An emoji-cloud was created showing the top 50 emojis in the dataset by absolute frequency, scaled proportionally to their frequency of use (see Figure 6). Many of the most frequently used emojis, like the face with tears of joy, heart, and face with heart eyes emoji, do not lend much insight into the thematic content of the dataset, so an accompanying word-cloud was also generated to visualize the most frequently used hashtags also scaled by frequency of use (see Figure 7).



Figure 6: The top 50 most frequently used emojis across the entire dataset. Emoji size is proportionate to the frequency of use of each emoji.

While some hashtags, such as #coronavirus and #covid19 seem to indicate popular topics of discussion, one surprising finding from the hashtag visualization shown in Figure 7 is the frequency of the p2000 hashtag. To investigate the usage of this hashtag, the Twitter hashtag explorer (www.twitter.com/explore) was used to find posts from 2020 containing this hashtag. Upon further research, #p2000 appears to be used primarily by a Dutch emergency alert system on Twitter. The account appears to be a non-human user that publishes tweets each time an emergency service such as an ambulance, fire truck, or police vehicle is deployed in the Netherlands. Since these tweets are published as often as every 15 minutes and each tweet contains the p2000 hashtag, this hashtag has an extremely high absolute frequency of 80,209. The same alert system is also responsible for the frequency of #ambulance and many hashtags that are abbreviations for Dutch metropolitan areas. This revelation brings to the forefront the influence of overactive, non-human users often referred to simply as "bots". The hashtag #wetter, for example, also appears to have a high absolute frequency (11,889 uses) due to a German bot that posts updates on weather conditions in Germany.

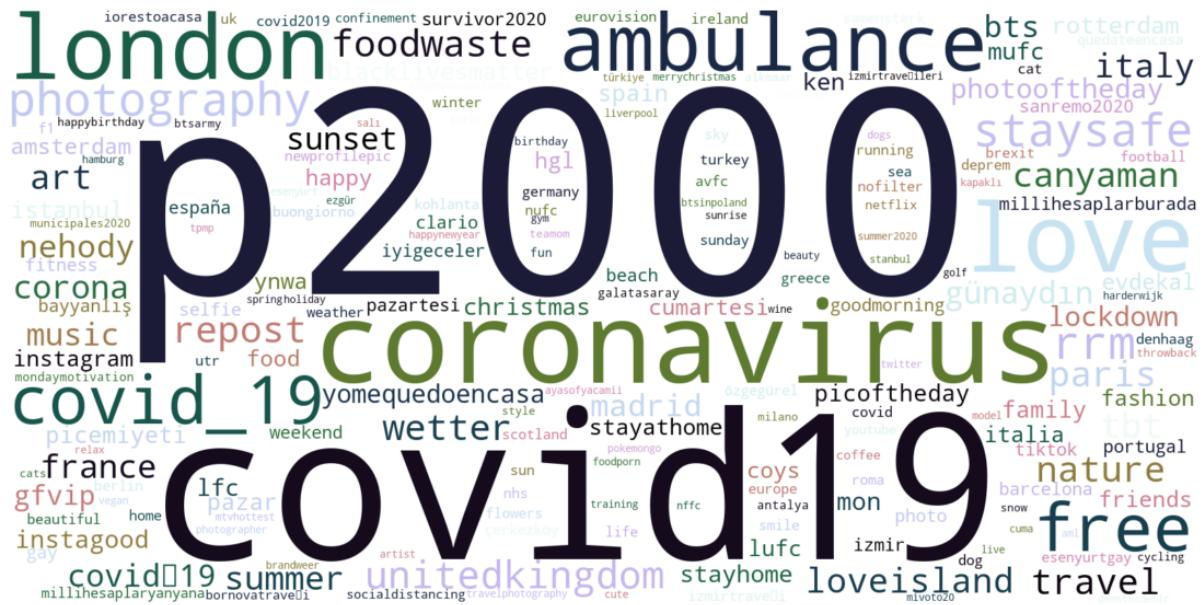


Figure 7: A word-cloud consisting of the top 150 hashtags used across the entire dataset. Text size is proportionate to the frequency of use of each hashtag.

Once some overall trends in emoji and hashtags usage were identified, more specific trends over time and space could be explored. In order to find out whether the usage of emojis changes over time and space, Tables 1 and 2 were prepared that show the top emojis by absolute frequency over time and space, respectively. The countries selected for the spatial table reflect the top 10 countries by user days (methodology explained in Section 3.3.3).

January	February	March	April	May	June	July	August	September	October	December
😊	❤️	😅	😂	❤️	❤️	❤️	❤️	😄	😄	❤️
❤️	😊	❤️	❤️	😊	😊	😊	😊	❤️	❤️	🎄
👑	👑	👑	👑	👑	👑	👑	👑	🔴	🔴	👑
🔴	🔴	🔴	🟡	🔴	🔴	🔴	🔴	👑	👑	👑
🎲	❤️	🟡	🚀	❤️	❤️	❤️	❤️	❤️	❤️	🔴
❤️	🎲	🤣	🤣	🤣	🤣	🔥	🔥	🤣	🤣	🤣
🔥	🔥	🏆	❤️	😊	😊	😎	😎	🔥	🔥	❤️
👉	👉	❤️	👉	👉	👉	😎	🤣	👉	👉	🔥
😊	😊	👉	😊	😊	🔥	😊	☀️	😊	😊	😊
😍	🖤	😊	🔴	🖤	🔥	👉	😊	😍	😍	🖤

Table 1: The most frequently used emojis for each monthly subset of the data, excluding November due to lack of data. Emojis listed in descending order by occurrence number.

United Kingdom	Spain	France	Germany	Italy	Turkey	Netherlands	Belgium	Switzerland	Austria
😂	❤️	❤️	❤️	❤️	❤️	🔴	❤️	❤️	❤️
❤️	😎	😂	😂	😂	👉	🚓	😂	😂	😎
👍	💪	👑	☂️	👑	😂	🏎️	👑	👑	😊
💪	💪	👑	😍	💪	❤️	🏎️	😍	❤️	👍
😍	😂	👍	👍	❤️	😍	❤️	🔥	👍	📘
🤣	❤️	🔥	👑	🤣	💛	😂	😎	👑	💪
❤️	🤣	😍	😍	♥️	👉	👍	👍	😎	😍
😊	👉	💪	😎	🔥	😊	💪	💪	💪	😊
👑	👉	👑	💪	👑	🔥	😊	💪	☀️	😊
👏	🔥	🤣	😊	♥️	💋	😊	♥️	😊	❤️

Table 2: The most frequently used emojis in each of the ten countries with the highest number of user days. Emojis listed in descending order by occurrence number.

From Table 1, it is clear that the top 10 emojis by absolute frequency do not stay consistent from month to month, suggesting some changes in emoji use occur over the course of a year, although many of these changes appear to be subtle. Table 2, on the other hand, demonstrates that more variation in emoji usage exists across countries.

The influence of non-human users was also evident in these tables. The red circle emoji, which appears as one of the top 10 most frequently used emojis in every single month of in Table 1, is included in every tweet posted by the previously mentioned p2000 Dutch emergency alert system. Therefore its presence in the table is most likely not indicative of its popularity amongst many people, but rather the influence of a single non-human, overactive user. This alert system also frequently uses the ambulance, police car, and fire truck emojis, which artificially inflates their popularity in the Netherlands as shown in Table 2. Similarly, the frequency of use of the umbrella emoji in Germany is not due to widespread user preference, but rather the result of a German weather update bot.

At this point in the analysis, it became clear that absolute frequency could only give surface-level insights into trends in emoji usage and that these insights would be subject to significant influence from hyperactive and non-human users. Simply converting from absolute to relative frequency was insufficient for the purposes of this research. The differences in emoji in Tables 1 and 2 support the claim that differences in emoji usage exist over time and space, but additional measures needed to be taken to gain deeper insights, reduce the influence of bots, and to look beyond the most frequently used emojis in each subset.

The typicality measure as presented by Hauthal et al. (2021) was implemented to address these concerns and gain a better understanding for where and when certain emojis typically occur. Since typicality values are normalized and calculated for individual emojis within a subset, it not only reduced the influence of overactive users, but also accounted for the varying amounts of data available for each monthly subset given

the temporal gaps in the dataset. Although other normalization and relative frequency measures exist, typicality was selected for use in this study due to its relative simplicity of calculation.

3.2.3 Temporal Typicality of Popular Emojis

Temporal typicality in the context of this research refers to the iterative calculation of each emoji's typicality for each monthly subset of the larger dataset. To avoid a common misunderstanding of typicality, it should be noted at this point that, due to the nature of the typicality measure, no list of "most typical emojis" for the whole given dataset can be calculated. Indeed, the calculation of typicality requires the data to be broken down into some sort of subset. Due to time constraints, it was not feasible to calculate typicality over time for every single available emoji. Therefore, other metrics such as absolute frequency and user days had to be used in order to prioritize which emojis are selected for processing and analysis.

The calculation of temporal typicality was therefore conducted for both the list of top 50 emojis by absolute frequency and the list of top 50 emojis by user days (methodology for the calculation of top emojis by user days described in Section 3.3.1). The full list of top 50 emojis by absolute frequency and by user days can be found in the Scripts folder of the Supplementary Materials in Raw_DataAnalysis.ipynb and HLL_DataAnalysis.ipynb, respectively (see Appendix E). These lists each have five emojis that do not occur in the other list. The victory hand (✌️), police car (🚓), oncoming fist (👉), hot beverage (☕), and white circle (⚪) emojis are included in the top 50 emojis by absolute frequency but not in the top 50 emojis by user days. Conversely, the face screaming in fear (😱), face with medical mask (😷), sun (☀️), rainbow (🌈), and winking face with tongue (😜) emojis are included in the top 50 emojis by user days but not in the top 50 emojis by absolute frequency.

For each list, the typicality of each emoji was calculated using each of the eleven available months as subsets (November excluded) and plotted as a line graph. Matrices of the results are shown in Appendices B.1 and B.2.

Some emojis remain fairly consistent over time, like the raising hands (🙌), ok-sign (👌), red heart (❤️), and thumbs up (👍) emojis, while others display significant variation over time, like the clapping hands emoji (👏) (which is typical for March and April), folded hands emoji (🙏) (also typical in March and April), sun emoji (☀️) (typical in June, July, and August), black heart emoji (🖤) (typical in June), party popper emoji (🎉) (typical in January and December), rainbow emoji (🌈) (typical in May and June), masked face emoji (🎭) (typical in March, April, and May), and the sun with face emoji (🌞) (typical in August and September). The degree to which emoji use changes over time depends on the emoji and no conclusions can be made about changes in general emoji use over time.

3.2.4 Spatial Typicality

After calculating typicality based on temporal subsets, spatial subsets were also used to identify spatial trends in emoji usage. Spatial typicality was conducted at two granularities: one analysis using country boundaries and another using 100 by 100 kilometer grid cells.

First, the spatial typicality analysis was conducted using country boundaries to denote spatial subsets. To do this, spatial joins were conducted for each of the ten selected countries to create subsets of data occurring within each country's boundaries. These countries were selected because they have the highest number of user days as determined during the HLL analysis (see Section 3.3.3). Then, a list of the 10 most frequently used emojis within each country was calculated, and the typicality of each emoji in the list was calculated using the country dataset as the subset. Many of the top 10 most frequently used emojis in each country demonstrated typicalities ranging from only -0.5 to 0.5. This would indicate that, especially for very frequently used emojis like the red heart (❤️) and laughing crying emoji (😂), they are so consistently popular across space that they are not typically found in any one location.

For instances of highly typical emojis occurring within the selected countries, a list of co-occurring hashtags was returned in order to determine the most common topics associated with that emoji. One notable instance of highly typical emojis occurs in the Czech Republic, where the automobile (🚗), vertical traffic light (⚡), and construction sign (🛠) emojis demonstrate typicalities of 165.05, 106.916372, and 143.16 respectively using posts geotagged within the Czech Republic as the subset for calculations. Upon further analysis, the top co-occurring hashtag by far for each of these three emojis is #nehody, the Czech word for accident. According to results from the Twitter hashtag explorer, this hashtag is used by a Czech emergency alert system quite similar to the #p2000 Dutch emergency alert system mentioned in Section 3.2.2. This alert system appears to be another hyperactive, non-human user that is one of the only users in the Czech Republic responsible for the use of the automobile, vertical traffic light, and construction sign emojis. This example illustrates the importance of further analysis for emojis that demonstrate high typicalities. Investigating the topical facet of the data by combining emojis and hashtags helps to reveal insights that might not be obvious just from statistical analysis.

Next, the most typical emojis were calculated for the ten countries with the most user days as calculated in Section 3.3.3. Using the existing subsets of data denoted by country boundaries, a function was created to calculate typicality for all emojis used at least 1000 times in each country. The threshold of 1000 uses was necessary to add because typicality values can become skewed when calculated for infrequently used emojis. This function returned a dictionary containing country names (keys) and dataframes with emojis in their generic form, the total occurrences of the emoji in the given subset, the name of the emoji, and the typicality of the emoji where the country dataset is used as the subset (values). The outputs of this function can be found in the Results folder of the Supplementary Materials (see Appendix E). Each of the resulting

dataframes was then filtered and ranked so that only emojis with positive typicality values within each country were included in the dataframe. The resulting dictionary was then exported for use during the emoji-specific analysis portion of the workflow. This method revealed several emojis with high typicalities that, due to lower absolute frequencies, would otherwise not have been investigated in this analysis. Emojis with positive typicalities for each country were further analyzed during the emoji-specific analysis in Section 3.4. For countries with many positively typical emojis, the top three most typical emojis were selected for analysis. In many cases, the emojis with the highest typicalities are some of the least frequently occurring emojis within the given threshold.

The next calculation of spatial typicality took place at a spatial data granularity of 100 by 100 kilometers and was visualized on a grid. Before the results of this analysis are analyzed, a short summary of the cartographic design process will be discussed. The goal was to create legible maps with minimalistic design to facilitate the analysis of dozens of maps at a time. One map would be created per emoji at this stage of analysis. The implementation of a user study to test map design and comprehension would certainly be beneficial to verify the legibility of the map design but was outside the scope of this thesis. Design decisions were therefore made and assessed with as much objectivity as possible based on basic cartographic design principles of legibility and compared with existing examples wherever possible.

For each of the 100 selected emojis, a choropleth map was generated by assigning each grid cell a color according to the typicality value of the emoji in that location. A custom color ramp ranging from beige to white to blue was used to represent negative, negligible, and positive typicality values, respectively. Beige was selected to represent negative values so that locations where an emoji is atypical would still be visible but less visually dominant than locations with strong typicality. The color palette diverges around white so that typicality values close to zero, which give us very little information about the typicality of an emoji at that location, are granted the least visual weight.

The size of grid cells used for visualization was also an important element in the design of these spatial typicality maps. For experimentation purposes, two data granularities were tested for visualization, one with grid cells measuring 50 by 50 kilometers and one with grid cells measuring 100 by 100 kilometers (see Figure 8). While the 50 by 50 kilometer grid allows for the more precise identification of local trends in emoji usage, the presentation of so much visual information at a fine granularity is not ideal for the efficient interpretation of spatial trends. The 100 by 100 kilometer grid, on the other hand, presents a smoother representation of spatial trends in the data and is quickly and easily legible - an important quality when interpreting results across one hundred emojis as was performed in this analysis.

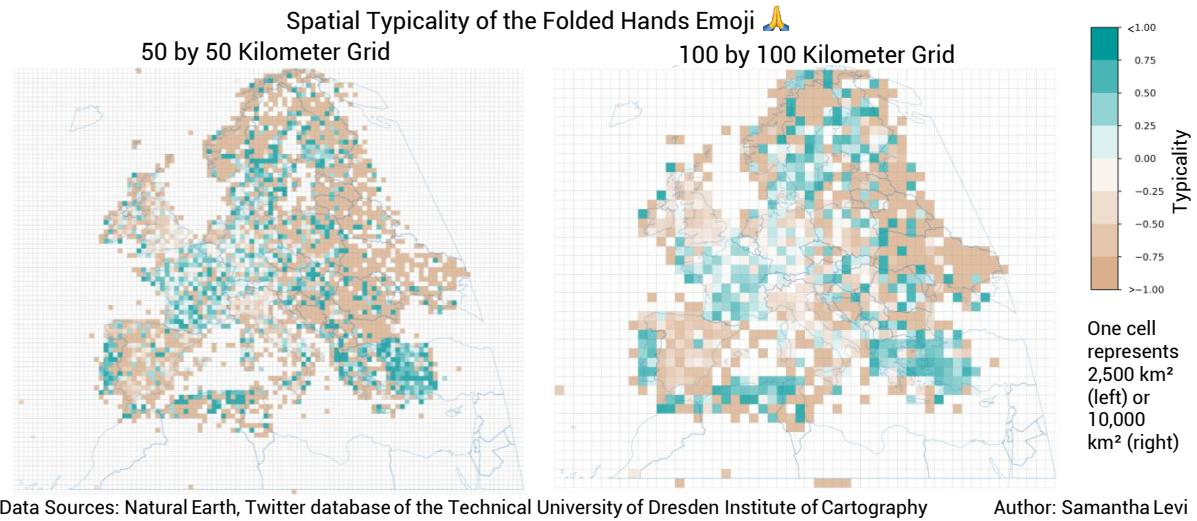


Figure 8: A comparison of the spatial typicality of the folded hands emoji (🙏) visualized on a 50 by 50 kilometer grid (left) and a 100 by 100 kilometer grid (right).

Other reasons for the selection of the larger grid include privacy awareness and the reduction of computational effort. The 100 by 100 kilometer grid cells offer a more conservative protection of location information (Dunkel et al., 2020). This is especially useful in this application because, unlike the data in the HLL format, the raw data used for spatial typicality calculations was not spatially aggregated during collection. Visualizing the data at a coarser spatial resolution allows for a more privacy-aware display of coordinate points. Since the spatial extent of grid cells is also used for the creation of the subsets used in typicality calculations, it is also beneficial to use larger grid cells since they produce similar visual results while requiring only one-quarter of the typicality calculations. This greatly reduced the processing time required to create each map.

Another critical element of the cartographic design was the decision not to crop grid cells to country boundaries. Cropped maps were created as part of the iterative design process using a simplified version of the country boundary shapefile to mask typicality values that do not occur on land (see Figure 9). However, when the results of the cropped and uncropped versions of the same emojis were compared, it was found that the cropped version, while having slightly improved aesthetics, obscured coastal data, in some cases preventing the visual detection of high-typicality areas. Because spatial data was aggregated, many data points in coastal areas were aggregated into cells that overlap with water bodies. By cropping out the portions of these grid cells that overlap water bodies, much of the data was obscured from final visualizations, thereby preventing accurate interpretations of emoji use across space. For example, in the cropped map representing the spatial typicality of the microbe emoji (🦠), regions of typical emoji usage in western Turkey and the Baltic Sea are difficult to interpret due to obscured coastal data. In the uncropped map, these regions of typical emoji usage are just as visible as other regions. Spatial context in the uncropped map is provided by the addition of country boundaries overlaid on the grid cells.

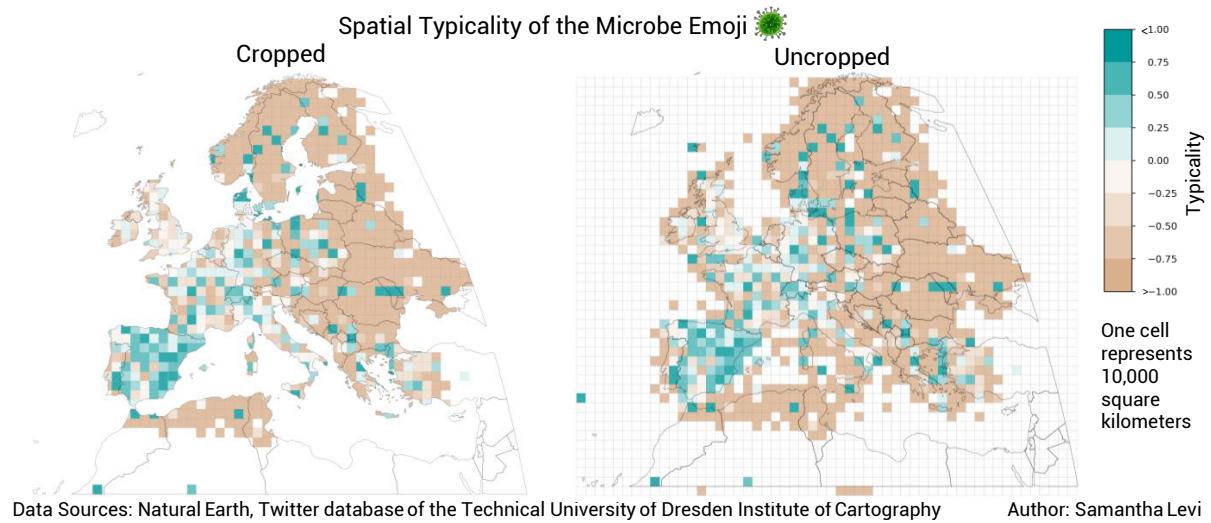


Figure 9: A comparison of the spatial typicality of the microbe emoji (_microbe emoji) in cropped (left) and uncropped (right) formats.

Here it should also be acknowledged that some grid cells in each map indicate typicality values over bodies of water. This phenomenon is documented in other visualizations of geo-social media data (Kejriwal et al., 2021) and is sometimes the result of tweets being published from small islands and boats. It is unlikely that these anomalies are due to imprecision during geotagging, since Twitter geotagging is both automated (thereby reducing the risk of data error from user specification) and precise to a ten thousandth of a degree (Malik et al., 2021).

Given these design considerations, the spatial typicality analysis was conducted at a spatial resolution of 100 by 100 kilometers (see Figure 10). This workflow allows for the visualization of emoji use over space to supplement the findings of the country-based calculations. The workflow of this section can be found in the notebook named SpatialTypicality_Grids (see Appendix E).

With a smaller dataset, a typical GIS workflow for grouping data points by polygon would be to conduct a spatial join between the point shapefile and the grid shapefile. However, this methodology would take an undesirable and infeasible amount of time and computational power when used on a dataset of over 4 million points. To circumvent this issue, a function was created using the Python NumPy library to find the best grid cell, or bin, for a given latitude and longitude input. Each tweet in the dataset was thereby assigned a corresponding row (y-match) and column (x-match) for the grid cell in which it was published. These values were then used to find a common index - in other words, to find all the grid cells in the generated grid that contained points from the dataset. It is worth noting that, due to the nature of this sorting mechanism, only rectangular grid cells were suitable for the analysis.

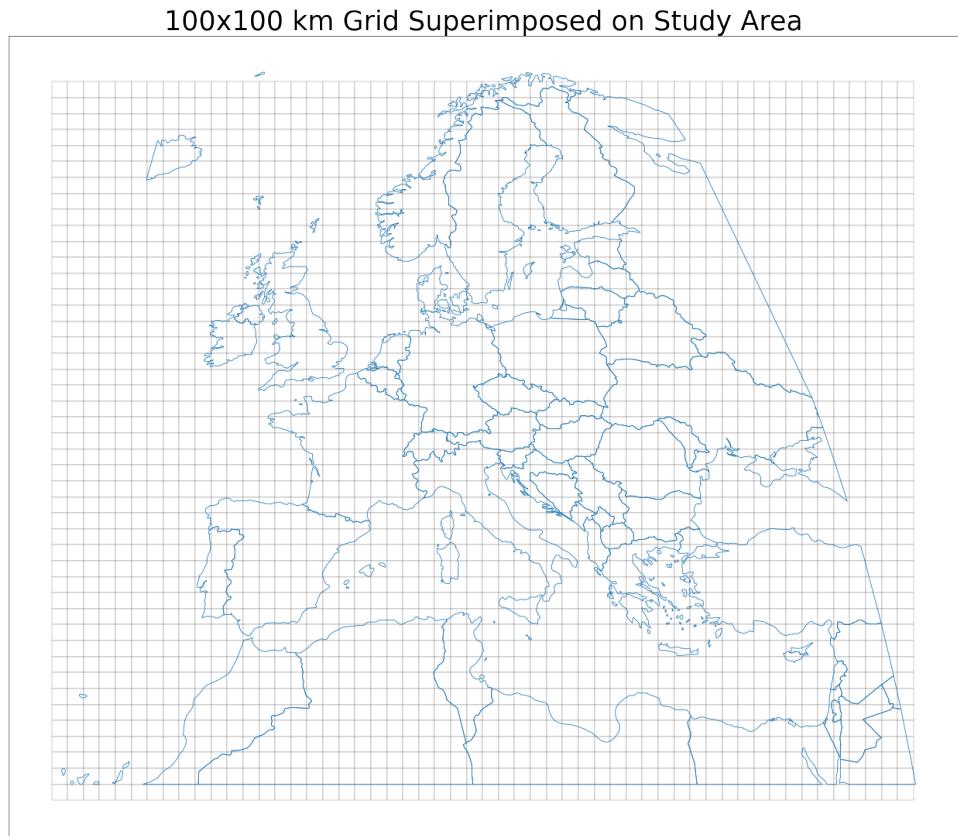
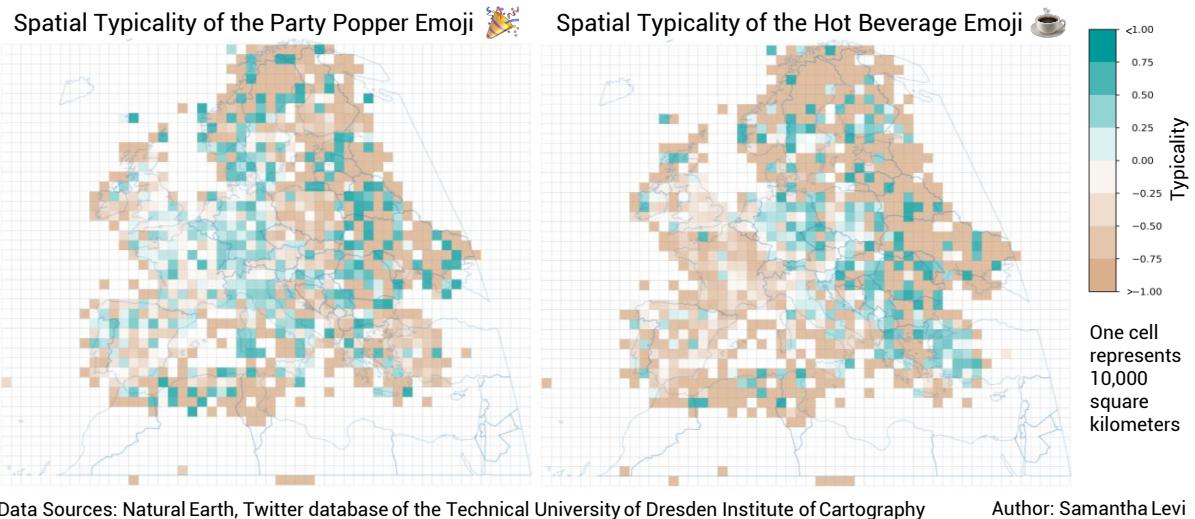


Figure 10: The grid generated for spatial analysis and visualization. Each cell represents an area of 10,000 square kilometers.

Once each point in the dataset was assigned to a corresponding grid cell, spatial typicality was calculated for the top 100 emojis by absolute frequency. That is, for each cell in the generated grid, the typicality of an individual emoji was calculated using all points assigned to the same grid cell as the subset. Once the typicality of each emoji within each grid cell was calculated, resulting maps were generated using the GeoPandas and Matplotlib Python libraries.

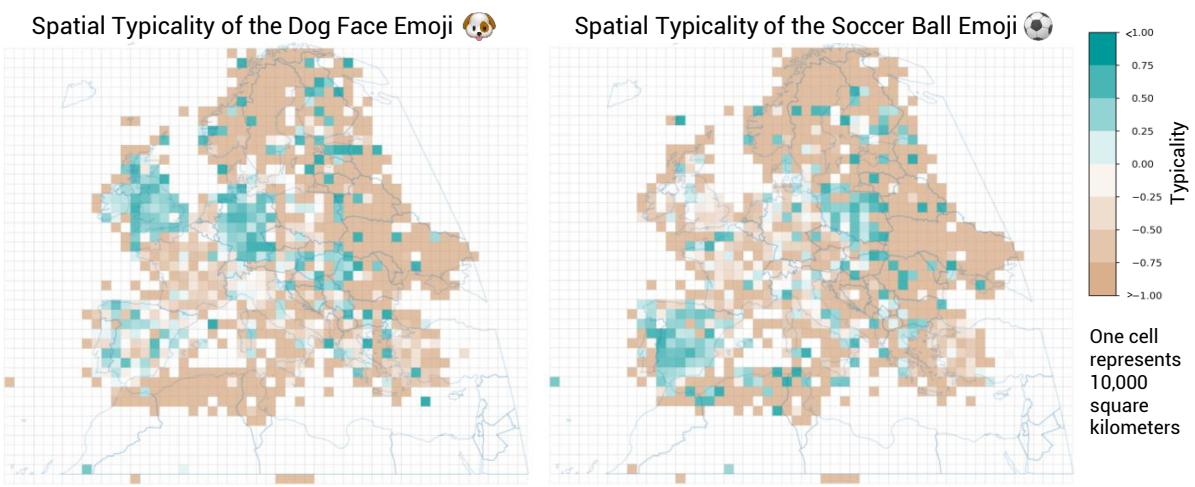
Each of the 100 resulting maps serve to visualize locations in which each emoji was found to typically occur. These maps effectively visualize and communicate the regional popularity of certain emojis. All of the resulting maps are available in the results folder of the Supplementary Materials repository under SpatialTypicality_Grids (see Appendix E). While some emojis like the party popper (🎉) emoji and the hot beverage (☕) emoji seem to demonstrate more ambiguous spatial trends (see Figure 11), some emojis display strong trends in use over space. Of emojis that demonstrate distinctive spatial patterns, some emojis seem to be typically used within specific countries, like the dog face (🐶) emoji in Germany and the United Kingdom and the soccer ball (⚽) emoji in Spain and Poland (see Figure 12). Others demonstrate geography-based spatial trends, like the water wave (🌊) which has positive typicality along the coast and the snowflake (❄️) emoji, which is typical along mountain ranges like the Alps and the Pyrenees as well as in Scandinavia (see Figure 13).



Data Sources: Natural Earth, Twitter database of the Technical University of Dresden Institute of Cartography

Author: Samantha Levi

Figure 11: The spatial typicality of the party popper (🎉) and hot beverage (☕) emojis. Both of these emojis demonstrate ambiguous spatial trends that do not give concrete insights into the use of either emoji.



Data Sources: Natural Earth, Twitter database of the Technical University of Dresden Institute of Cartography

Author: Samantha Levi

Figure 12: The spatial typicality of the dog face (🐶) and soccer ball (⚽) emojis. Both of these emojis demonstrate distinctive, country-based spatial trends.

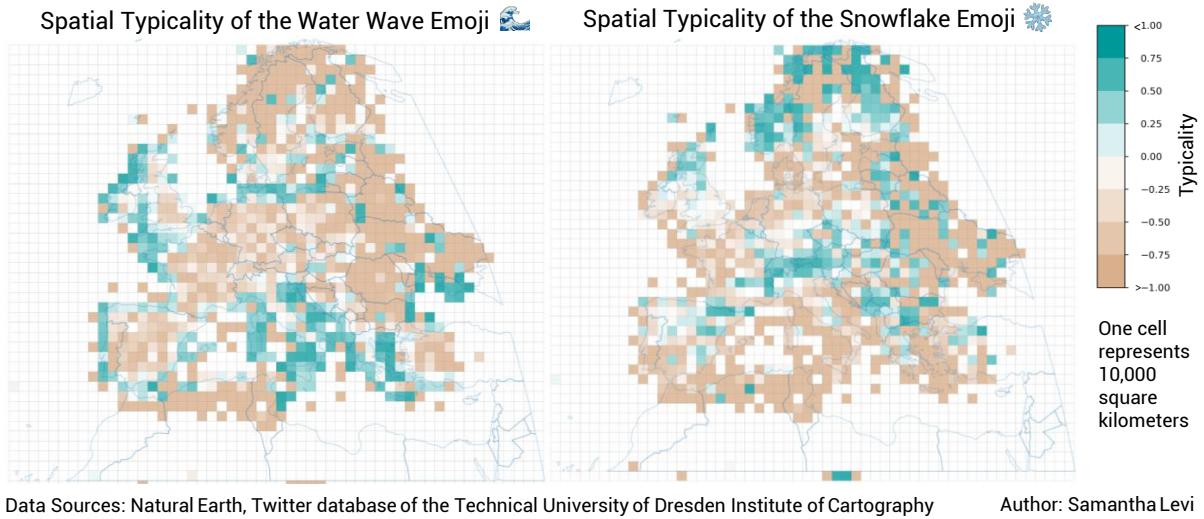


Figure 13: The spatial typicality of the water wave (浪) and snowflake (❄️) emojis. Both of these emojis demonstrate distinctive, environment-based spatial trends.

3.3 HLL Data Analysis

While the raw data format was sufficient for the investigation of the topical, spatial, and temporal facets of the data, privacy concerns prevented its use for the investigation of the social facet, since this would require the storing of sensitive information like usernames alongside precise location information. Although the information contained in Twitter posts can be considered volunteered information due to the public nature of social media platforms (Dunkel et al., 2020), it is still in the interest of this study to reduce the sensitivity of the dataset.

To address some of these privacy issues, the HyperLogLog data format was implemented alongside other data abstraction measures like cryptographic hashing and spatial data aggregation. The main advantage of HLL data over raw data is that the HLL format allows for the social facet of the data to be explored without storing sensitive user information. For example, counting the number of distinct days on which individual users published a tweet using the raw data format would require storing both user IDs and date information, which would potentially allow each user to be tracked across space and time. Counting user days with HLL data does not require this information to be stored, therefore protecting users from being tracked. Because the HLL data was previously generated based on the original data, qualifications about data completeness from the raw dataset also apply to the data in the HLL format.

As mentioned in Section 3.1.1, multiple aggregation levels of HLL data were possible, from which a data aggregation level of 4 was selected (see Figure 2). While some accuracy is sacrificed with these measures, the result is a dataset with an acceptable, conservative privacy-accuracy trade-off that allows for data analysis while offering users not insignificant protection from malicious entities. As discussed in Dunkel et al. (2020), data in the HLL format is only able to be compromised in the event of an intersection attack under very unique and improbable circumstances.

The HLL format also permits three important calculations: union, intersection, and cardinality. The lossless union and intersection functions are both set-based functions that calculate the number of distinct items in a merged dataset and in independent datasets, respectively. Cardinality, as mentioned earlier, is the number of unique items in a set. It can be used to calculate three different user-based metrics: user count, post count, and user days, as demonstrated in Wood, Guerry, Silver, and Lacayo (2013), Dunkel et al. (2020), and Y. Kim et al. (2019). The term user day in this case is borrowed from Wood et al. (2013) and in the context of this research refers to the total number of days, across all users, that each user published a tweet within the study area.

To illustrate the necessity of implementing the cardinality function rather than adding together multiple user day column values, consider a hypothetical post containing the text "Great day today! #summer 🌞☀️👍😊👍😊👍😊". This tweet would show up as 6 different rows in the HLL dataset, all with the same location information in the latitude and longitude columns but with different versions of the thumbs-up emoji in the emoji column. Each row would have a user day value of 1. Adding together the user days column over these 6 rows would therefore incorrectly return a value of 6 user days. The correct cardinality of user days in this case would be one, because each row refers to a single post created by a single user on only one day. To achieve the correct result, the cardinality of user days needed to be recalculated for all posts containing each skin tone variation of the thumbs-up emoji. To do this, multiple subsets were created for posts containing each emoji variant. A union function was then performed in order to create a single subset representing all usages of the thumbs-up emoji. Finally, the cardinality function (Appendix C.1) could be implemented to calculate the approximate number of user days for the thumbs-up emoji.

Using this workflow, unions were performed on subsets of the data containing similar emojis with different skin tone modifiers to allow for accurate calculation of post count, user count, and user days. The Python library `python-hll` was used for purposes of HLL implementation. To illustrate the necessity of performing the union and implementing the cardinality function, consider the following example: the number of total user days in France was 390,921 when calculated as a sum of user day column values, and only 228,942 when calculated as the cardinality of user days for the union of all posts geotagged in France. Even assuming the average error rate of 2% for HLL cardinality estimations (Dunkel et al., 2020), using the cardinality function avoids an overestimation of 157,401 user days in France.

Unfortunately the nature of the HLL format made it impossible to analyze all four facets of LBSN at once. The data on the remote server is structured into tables separated by facet and therefore no table existed in which hashtags, emojis, coordinates, and user days could be accessed at once. The construction of a new HLL dataset converted directly from raw data requires considerable time and labor and was outside of the scope of this project. Ultimately, an HLL dataset without hashtags or publishing date was used, meaning that the temporal and topical facets could not be properly analyzed using HLL data alone. Raw data was therefore used to investigate the temporal, spatial,

and topical facets of the geo-social media data, while the HLL data was used to explore the social and spatial facets of emoji occurrence. Since neither dataset was sufficient for the full exploration of the dataset, the results of both data analyses will later be considered in tandem.

3.3.1 Calculation of User Days per Emoji

Once the HLL data was collected from the remote server and read into the notebook as a geodataframe, the first task was to remove the skin tone modifiers from the emojis so that similar emojis could be counted together, as in the raw data preprocessing. In contrast to the methodology used with the raw data (discussed in Section 3.1.2, the union and cardinality functions of the HLL data format has to be leveraged in order to estimate the number of user days per emoji.

Using the union function shown in Appendix C.2, unions were performed on emojis with potential for skin tone modification to allow for accurate calculation of user days as shown in Figure 14. To reduce computational effort, this was only performed for emojis with skin tone modifiers that appeared in the list of top 100 emojis by user days. This step results in significant differences in the number of user days calculated. For example, the generic thumbs up emoji was calculated as having 70,026 user days, while the thumbs up emoji including all of its variations was calculated as having 95,117 user days.

```

# make dictionary of emoji names and their variations
emojidictemo = {
    ":clapping_hands": "\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d",
    ":folded_hands": "\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d",
    ":thumbs_up": "\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d",
    ":flexed_biceps": "\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d",
    ":OK_hand": "\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d",
    ":raising_hands": "\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d",
    ":backhand_index_pointing_down": "\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d",
    ":backhand_index_pointing_right": "\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d",
    ":victory_hand": "\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d",
    ":oncoming_fist": "\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d\ud83d\udc4d"
}

emoji_ud = {}

for name, variations in emojidictemo.items():
    subset = gdf[gdf['emoji'].str.contains(variations)]
    emoji_ud[name] = union_all_hll(subset["pud_hll"].dropna())

emoji_ud

{'clapping_hands': 99485,
 'folded_hands': 93921,
 'thumbs_up': 95117,
 'flexed_biceps': 87392,
 'OK_hand': 50397,
 'raising_hands': 45654,
 'backhand_index_pointing_down': 36136,
 'backhand_index_pointing_right': 26892,
 'victory_hand': 21772,
 'oncoming_fist': 24498}

```

Figure 14: Unions were performed on emojis with multiple possible skin tones. Then, the number of user days per generic emoji was calculated.

Once the unions were performed, the cardinality function could be implemented to calculate the approximate number of user days for each emoji. The cardinality function is shown in Appendix C.1. Due to time constraints, it was not possible to calculate the number of user days for every single emoji; the list had to be cut down somewhat. Therefore a list of top 100 emojis by user days was created in which generic and skin-tone-specific variations of emojis would have the same number of user days. The cardinality of posts containing each emoji was also calculated for each emoji in that list and added to a separate dictionary.

Finally, a third dictionary was created containing the list of top 100 emojis and the difference between the respective post count and user days. This calculation was conducted in order to find which emojis have a large discrepancy between post count and user days. Emojis with a larger discrepancy value are included in many posts by the same user on the same day. It is therefore possible to see from this difference dictionary

which emojis are most often used by overactive users. This can further provide insights as to which emojis are most often used by non-human overactive users such as bots. Indeed, the emoji with the greatest discrepancy is the red circle emoji, which was proved in the raw data exploration phase to almost exclusively be used in Dutch emergency alerts labeled with the hashtag #p2000. The police car, ambulance, and fire truck emojis, which are all also associated with the same emergency alert bot, also appear in this list at positions 5, 8, and 21, respectively.

3.3.2 Visualization of User Days

In order to gain a better understanding of the spatial distribution of the data without plotting the precise coordinates of each tweet, a grid of 100 by 100 kilometer cells was used to visualize the number of user days across the study area. User days was selected as a metric as opposed to post count or user count in order to reduce the influence of overactive users (who would artificially drive up post counts) without counting non-active users (who would artificially drive up the user count). Other cell sizes, like 50x50 km and 200x200 km sizes, were also created for comparison. However, the 100x100 km cell grid was determined to be the best granularity for visualization purposes and was therefore used for later map production. Visualizing data at this coarse granularity also further protects the privacy of users' precise locations.

As discussed in Section 3.2.4, conducting a spatial join between post locations and the 100 by 100 kilometer grid would take undesirable amounts of computational effort and time. Therefore, during spatial data analysis, the function `get_best_bins` was once again implemented to assign each post to a corresponding grid cell given its coordinates. Suitable x and y bins that were then saved in the geodataframe as index columns. Although other studies that calculate user days have implemented non-rectangular grid cells for visualization purposes over a smaller study area (Y. Kim et al., 2019), the nature of this binning technique requires the use of square grid cells.

Once the grid was created and points were assigned to suitable x- and y-bins, a function was created that would run a union on all points within a grid cell and calculate the cardinality of user days contained within the cell. This function was used to iteratively calculate the cardinality of user days within each grid cell containing data and the outputs were saved as a geodataframe. The resulting map, shown in Figure 15, demonstrates the spatial distribution of user days in the study area. In other words, this map demonstrates the number of users who published at least one tweet within the given grid cell. The minimum legend classification includes values between 1 and 500 in order to avoid instances where individual users could be singled out and their information therefore potentially compromised.

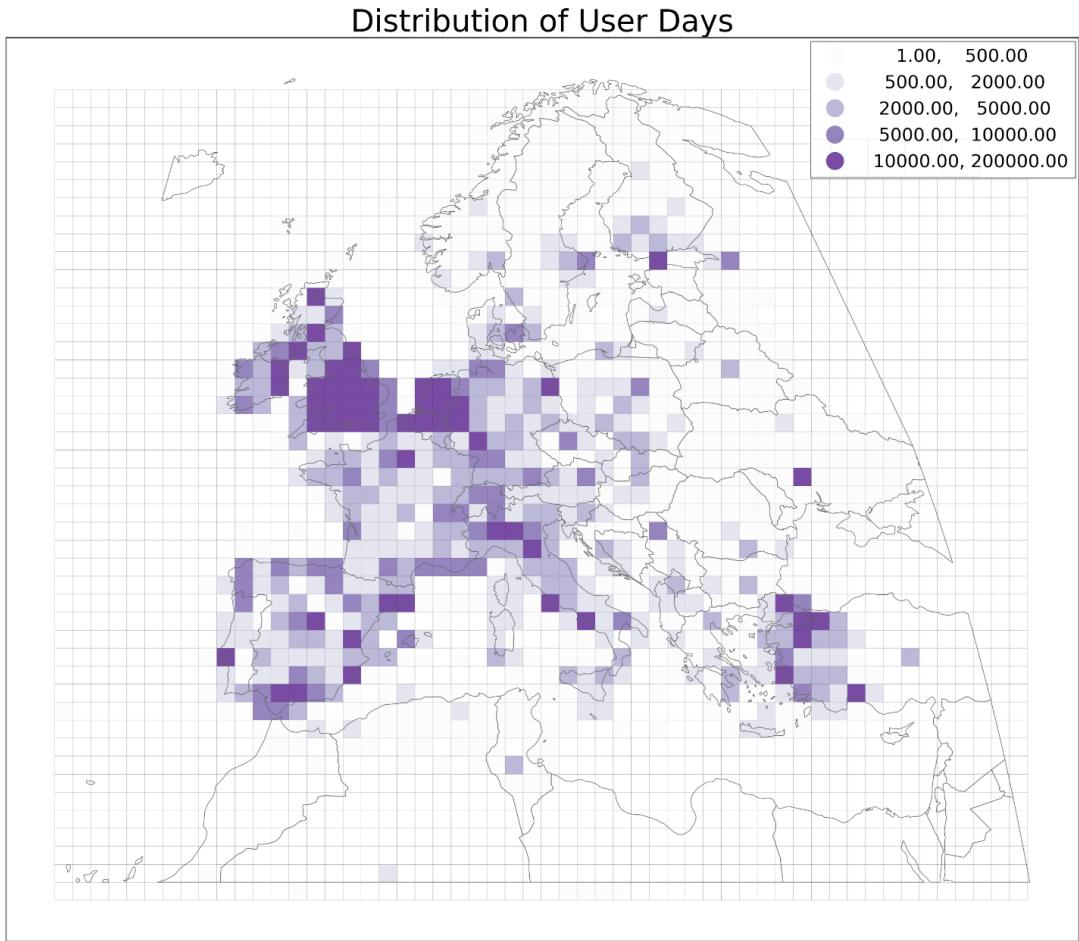


Figure 15: The number of user days (days where distinct users posted at least once) per 100 by 100 kilometer grid cell.

Rather than solely visualizing the number of users or the number of posts that occur within each grid cell, the user day metric shown in Figure 15 gives an insight into the distribution of Twitter activity across the study area. If a given grid cell is assigned a value of 1000, for example, this means that approximately 1000 distinct users published a geotagged tweet including at least one emoji and one hashtag in that area over the course of a year. Darker areas in the map therefore represent hotspots of user activity that mitigate the influence of hyperactive users. Since time limitations restricted the number of countries for which user day calculations could be conducted, these hotspots were also used to inform the selection of relevant countries for user day analysis.

3.3.3 Calculation of User Days per Country

Another calculation that was conducted with HLL data was the cardinality of user days by country. This step was deemed essential after the initial results of the raw data exploratory analysis demonstrated the need to narrow down the scope of analysis using a metric other than absolute or relative frequency. A list of the top 10 countries by user days was therefore generated in order to create a meaningful spatial subset of the data on which later workflows could be performed and tested.

To calculate the number of user days within each country, the 100 by 100 kilometer bins used in the previous section were used to follow the methodology demonstrated in Dunkel et al. (2020). Firstly, individual countries were extracted from a shapefile of country administrative boundaries. Grid cells were then selected whose centroids intersected the country geometry. The resulting cluster of grid cells representing each country was then spatially joined with the HLL data points. Finally, a union function was implemented to join together all of the now data-enriched grid cells and the cardinality of user days was calculated for the entire country. Figure 16 shows this process using the United Kingdom as an example. This process was conducted for 12 countries selected for their coverage within the dataset, and from these countries 10 were selected as having the highest number of user days.

```

uk = europe[europe['NAME_EN'] == 'United Kingdom']
# make country grid
grid_uk = intersect_grid_centroids(
    grid=grid, intersect_gdf=uk)
# join hll data to grid
uk_hll = gdf.sjoin(grid_uk, how="right")
# join together country grid cells, calculate the number of userdays
# within the cluster of grid cells
userdays_uk = union_all_hll(uk_hll["pud_hll"].dropna())
# add the result to a dictionary
country_ud["United Kingdom"] = userdays_uk

```

Figure 16: The calculation of user days within the United Kingdom.

Here it is important to note that the cardinality calculations are also somewhat dependent on aggregation level - when the user days per country was performed on HLL data with a finer spatial granularity (spatial aggregation level of 5), the results differed (see Table 3). This example illustrates how the Modifiable Areal Unit Problem (MAUP) affects the results of the analysis. For most countries, the differences between results are minimal, with the exception of Switzerland and, to a lesser extent, the Netherlands. Despite the effects of the MAUP, however, the ranking of the top 10 countries by user days remains consistent.

Country	User Days (Agg. Level 4)	User Days (Agg. Level 5)	Difference
United Kingdom	811956	810535	0.18%
Spain	288547	288819	-0.09%
France	228942	230294	-0.59%
Germany	143224	142974	0.17%
Italy	143012	141807	0.85%
Turkey	111138	108351	2.57%
Netherlands	76856	73083	5.16%
Belgium	40219	39852	0.92%
Switzerland	20624	23061	-10.57%
Austria	17732	18070	-1.87%
Portugal	13177	12699	3.76%
Czech Republic	10485	10711	-2.11%

Table 3: Differences in user day calculations due to the aggregation level at the time of data collection.

3.4 Emoji-Specific Analysis

After conducting the raw and HLL data exploratory analyses, it became clear that a distinction should be made between topic-specific and non-topic-specific emojis. For example, when looking at the results of the temporal typicality matrices (Appendices B.1 and B.2), certain emojis not only exhibit more distinctive variations in typicality over time, but also seem to represent more concrete topics. For example, the typicality of the face with medical mask emoji over time varies much more over time than the typicality of the red heart emoji. Between these two emojis, the face with medical mask emoji also seems to represent a much more specific concept (medical masks) than the red heart emoji (love).

The same trend is true for many of the emojis included in the temporal typicality matrices. Very frequently used emojis in particular may be popular exactly because their meanings are ambiguous and can be applied to many posts regardless of the topic being discussed. Additionally, only a few of the emojis analyzed seem to have significant changes in typicality over time. Most emojis analyzed demonstrate typicality values that remain consistently close to zero over time. This should not be interpreted as the lack of use of the particular emoji, but rather as the consistency of the emoji's use over time.

Because many of the emojis analyzed did not demonstrate significant change in typicality over time, the topical facet was not investigated for the full list of top 50 emojis by absolute frequency or user days. Even if, for example, the winking face emoji was found to be topic-specific, this result would do little to illustrate where or when that topic was being discussed. Therefore, due to the processing time required for each emoji and in order to avoid running copious amounts of analysis on non-topic specific emojis with potentially superficial results, the focus of analysis was restricted to emojis that

were either hypothesized to be topically consistent, or who were found to be particularly typical within one of the top 10 countries by user days. Once the topical facet of each emoji was established, spatial and temporal typicality analyses were conducted for topic-specific emojis. Results from selected emojis will be individually interpreted in Section 3.4.5.

3.4.1 Topical Consistency

Spikes in emoji typicality in certain locations or during certain times of the year lend themselves to speculation. One person might hypothesize that the positive typicality of the rainbow emoji (🌈) in June is due to users reacting to rainbows they see outside in early summer, while another might guess that the spike in typicality is due to the discussion of lesbian, gay, bisexual, transgender, queer, intersex, and asexual (LGBTQIA+) rights during June, which is international LGBTQIA+ pride month. While both of these hypotheses have logical reasoning behind them, additional research is necessary to prove whether the rainbow emoji and other emojis consistently refer to the same topic. The intention of this section of the workflow was to identify topically consistent emojis whose variable typicality over time and space could lend insights about popular topics of discussion on Twitter.

To establish the methodology, two emojis were selected that were hypothesized to be topically consistent: the wine emoji (🍷) and the beer mug emoji (🍺). These emojis were selected because they represent concrete objects that are similar to each other (both alcoholic beverages) but still distinguishable and specific enough that differences in the topical facet should still be detectable during analysis. A viable methodology for the determination of topical consistency will be able to successfully differentiate these two emojis.

In this analysis, the topical facet of the data is defined as the combination of emojis and hashtags that occur within the same post. Therefore, in order to investigate the topical facet of the given dataset, a list of 20 co-occurring hashtags was generated for each emoji selected for analysis. From this list, hashtags were identified that discussed similar topics, and the number of posts containing the each hashtag were quantified and divided by the total number of posts containing the top 20 co-occurring hashtags. The resulting percentage is the topical consistency of the given emoji. In order to determine what topic an individual hashtag is referring to, the Twitter hashtag explorer was used to gather related posts and compare them for topical consistency. The number of co-occurring hashtags in the list comes with a time-accuracy trade-off; more hashtags in the list could return more precise results but would also take more time to analyze. Since many hashtags in the list had to be individually contextualized and interpreted, it was not feasible within the time frame of the project to include more than 20 hashtags per emoji.

To visualize the topics associated with each emoji, the top twenty co-occurring hashtags per emoji were assembled into word-clouds using the WordCloud Python library.

Each hashtag in the word-cloud is proportionately scaled according to its frequency of occurrence within posts containing the associated emoji. These visualizations allow for a fast, general understanding of the subjects most frequently associated with a given emoji, as seen in Figure 17. This table reveals that the emoji was most commonly used in tweets containing hashtags referring to the COVID-19 pandemic and appreciation for healthcare workers.

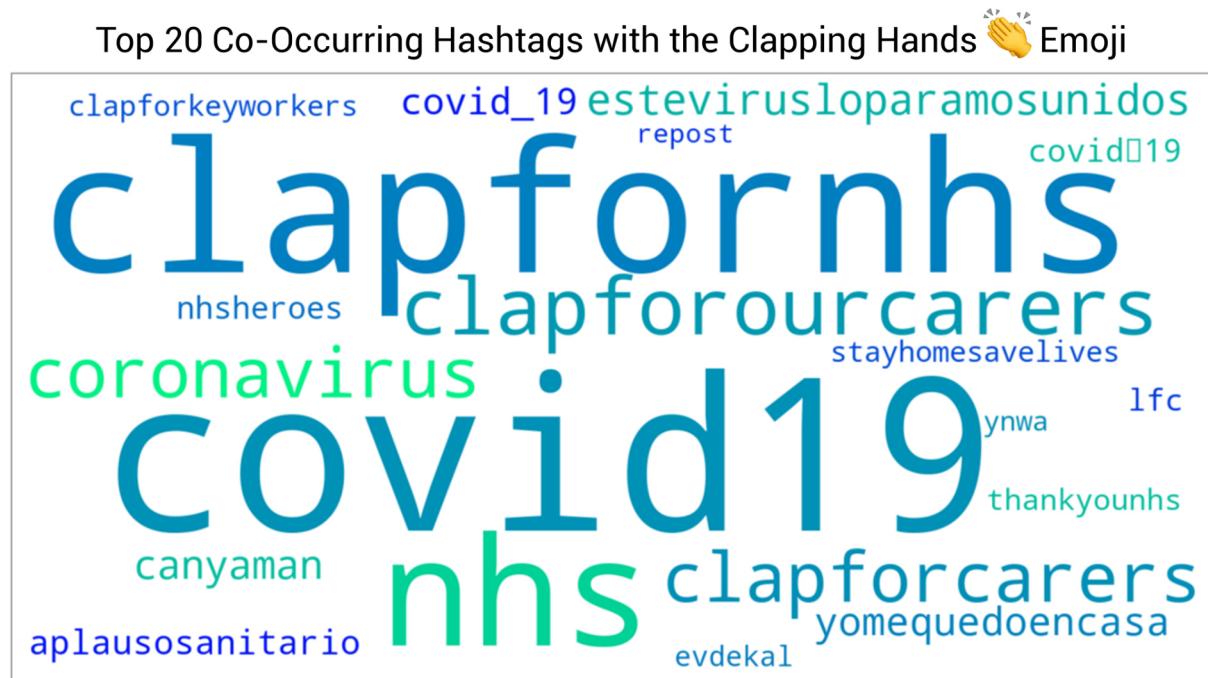


Figure 17: The top 20 co-occurring hashtags associated with the clapping hands (牚) emoji.

Since the beer mug and wine emojis both represent concrete objects, their percentages of topical consistency were used as a guideline to assess other emojis. The beer mug emoji demonstrated a topical consistency of 92.1% and the wine emoji demonstrated a topical consistency of 75.4%.

The lower percentage of topical consistency for the wine emoji is due to the conservative definition of topically consistent hashtags as only those hashtags that occur almost exclusively within posts discussing the given topic. In other words, co-occurring hashtags must be both topically consistent and topically specific. For example, the hashtags #france and #italy were both included in the list of top 20 co-occurring hashtags with the wine emoji. While both of these countries are commonly associated with their wine production in colloquial life, posts containing the hashtags #france and #italy could discuss any number of other topics associated with those countries. To avoid the over-estimation of topical consistency, both hashtags were therefore not considered to be topically specific to the wine emoji and were not included in the calculation of topical consistency.

Based on these benchmark calculations, an emoji was said to have topical consistency

for the purposes of this study if more than 70 percent of posts using the top 20 co-occurring hashtags refer to the same topic. To this end, occurrences of thematically similar hashtags were summed up and divided by the total number of posts created in each co-occurring hashtag table. An example of this methodology is shown in Table 4 for the raised fist emoji, which was confirmed as being symbolic of the Black Lives Matter movement. To see the co-occurring hashtag tables for all emojis analyzed, see the Supplementary Materials in Appendix E.

Raised Fist Emoji			
Rank	Hashtag	Uses	Calculation
20	blackoutuesday	93	Total Posts:
19	bluecollarstore	95	7343
18	blacklifematters	97	BLM Posts:
17	8m	98	6678
16	blacklivesmatteruk	98	% Consistency:
15	repost	101	90.94%
14	montmartre	107	
13	justicepouradama	113	
12	paris	114	
11	justiceforgeorgefloyd	123	
10	blacklivesmattters	134	
9	glasgow	150	
8	blackisking	156	
7	blacklivesmattter	159	
6	georgefloyd	189	
5	blm	506	
4	respect	577	
3	notoracism	688	
2	blackouttuesday	1128	
1	blacklivesmatter	2617	

Table 4: An example of a co-occurring hashtag table with topically consistent hashtags highlighted in gray. This process was repeated for each of the 35 emojis that were analyzed for topical consistency.

While some hashtags had relatively straightforward meanings, additional investigation was often necessary in order to determine topical consistency. For instances where hashtags were previously unknown to the author, the Twitter hashtag explorer was used as a way to "ground-truth" the data with real examples of how the hashtag was used. Thematically ambiguous hashtags, such as names of cities, were always considered to be non-topic specific, resulting in more conservative estimates of topical consistency. Hashtags not written in English were translated either by the author or using online translation services. In instances where hashtag meanings were still unclear even after

translation into English, the author was able to consult with native speakers to deduce their meaning in context.

Ultimately, 35 emojis were investigated for topical consistency. Ten of these were selected based upon their hypothesized connection to relevant topics, and the remaining 15 were analyzed because they were found in Section 3.2 to be typical in one of the top 10 countries ranked by total user days. For a complete summary of the results of this topical consistency workflow, see Appendix D.1.

3.4.2 Temporal Typicality

To supplement the information from the absolute frequency of each emoji's usage, a bar chart of typicality over time was also generated for each emoji using monthly subsets of the overall dataset. Although weekly or biweekly subsets could theoretically also be implemented for the calculation of typicality, monthly subsets were used in order to minimize the number of empty subsets. These charts are especially illustrative of the benefits of the typicality measure because of the temporal gaps in the data collection. For example, several weeks worth of data from April 2020 were not able to be collected, which leads to a somewhat misleading drop in overall emoji usage for the month of April (see Figure 18). If one were to only analyze the total number of emojis that were collected, one might incorrectly assume that equal number of tweets were collected for each available month and conclude that users used fewer emojis in April compared to March.

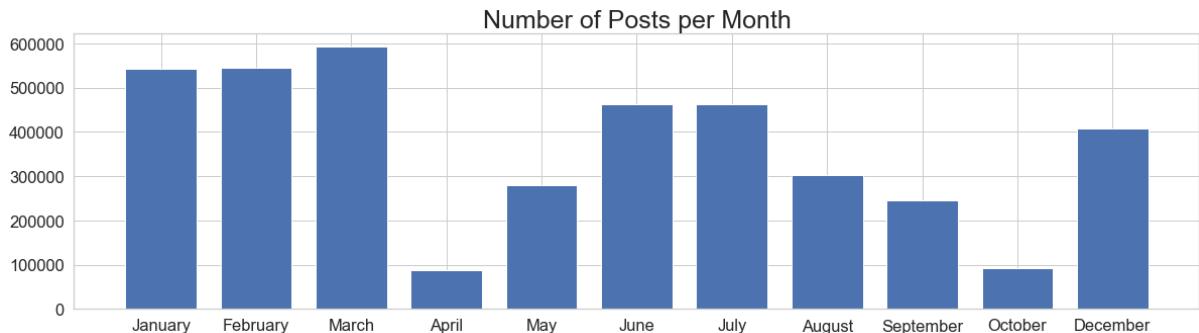


Figure 18: The number of tweets collected per month in the raw Twitter dataset. Some unavoidable gaps in data collection create inconsistencies from month to month. No data was collected for the month of November.

For example, given the bar chart of total usages of the face with medical mask emoji (😷) shown in Figure 19, one might incorrectly assume that the mask emoji was hardly used during the month of April. However, this drop in usage is actually due to the lack of available data for April, and is not the result of changing user behaviors. Because the typicality measure is normalized based on the size of the subset, it is ideal for investigating such instances. This distinction is clearly illustrated in the chart of temporal typicality for the face with medical mask emoji shown in Figure 20. Here it

is obvious that, although the total number of recorded emoji usages went down from March to April, the typicality remains positive.

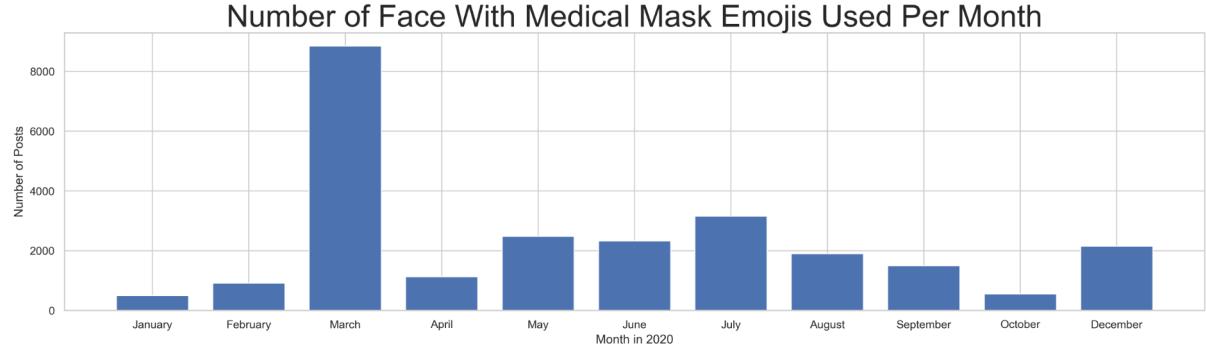


Figure 19: The absolute frequency of the face with medical mask emoji (😌). The significant drop after March is misleading due to a gap in the dataset during the month of April.

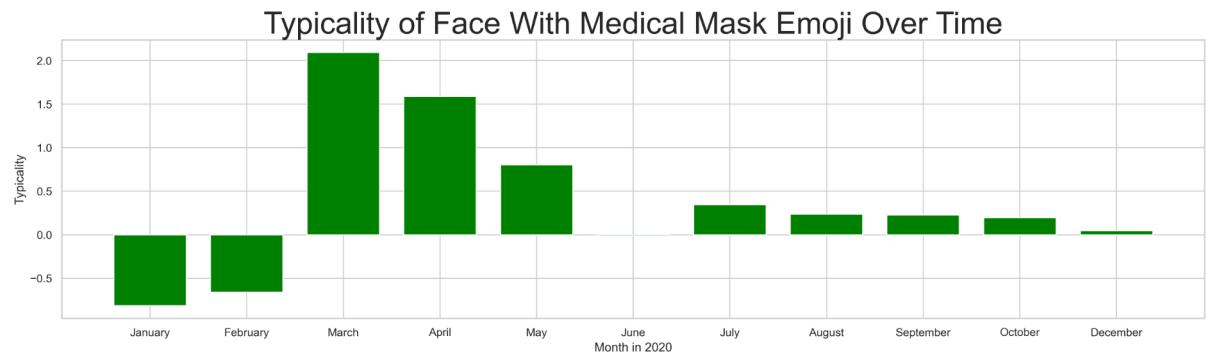
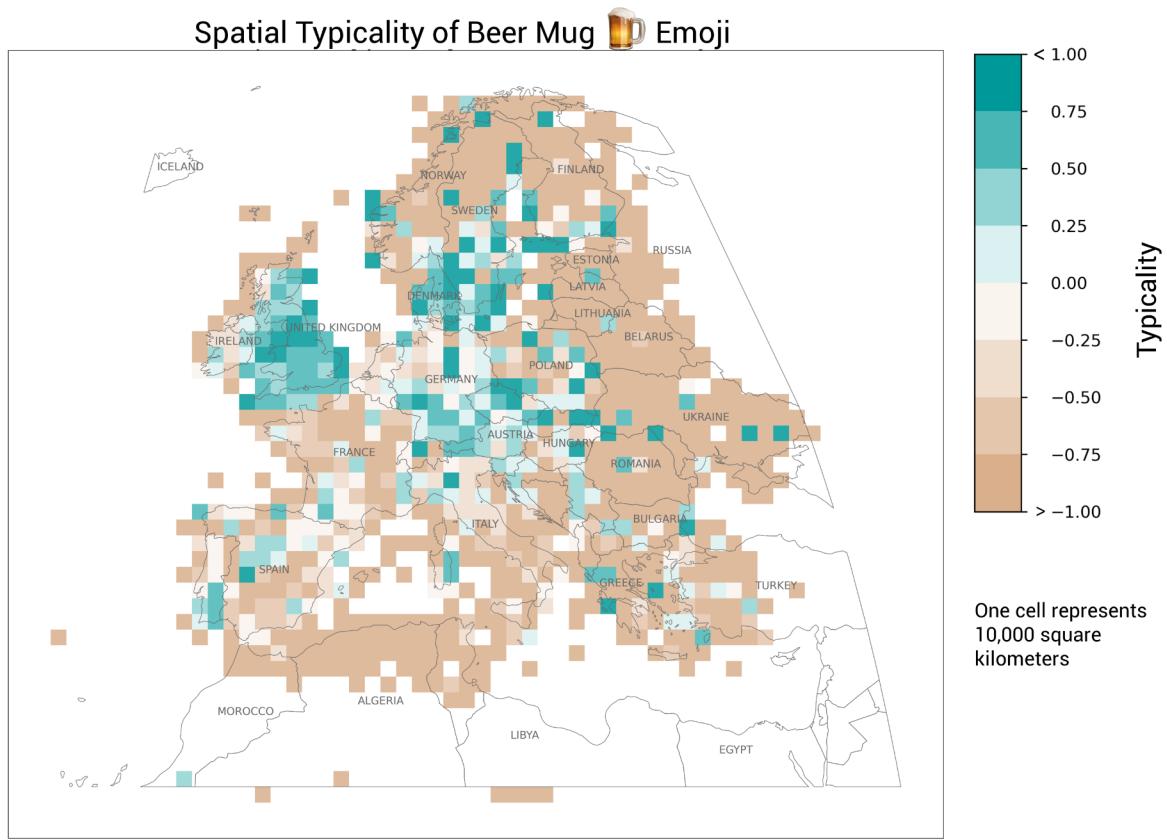


Figure 20: Typicality of the face with medical mask emoji (😌) calculated using monthly subsets.

3.4.3 Spatial Typicality

For each of the selected emojis, maps showing the typical locations of each emoji were produced following the methodology laid out in Section 3.2.4. To improve the aesthetics of these maps as well as to provide additional spatial context, the cartographic design of these maps was refined; grid cells were overlaid on top of a custom basemap consisting of country borders and labels to provide additional spatial context. In order to avoid visual clutter, only large European countries were labelled. One example of the output maps is shown below in Figure 21. Several iterations of this map style were generated and compared before the final visualizations were created. Among other things, several color ramps, text sizes, grid line thicknesses, country boundary thicknesses, and grid cell opacities were tested. The iterative design process can be found in the Jupyter Notebook Animation_CreateFrames_Visualizations.ipynb in the Supplementary Materials (Appendix E).



Data Sources: Natural Earth, Twitter database of the Technical University of Dresden Institute of Cartography

Author: Samantha Levi

Figure 21: Typicality of the beer mug emoji (🍺) calculated using 100 by 100 kilometer grid cells as spatial subsets.

These maps demonstrate the typicality of each emoji across the study area without taking into account the temporal facet. They give no insight into when the topic associated with the emoji might have been discussed, but rather show just where these discussions may have occurred. For example, in Figure 21, the beer mug emoji is shown to typically occur in the United Kingdom, Germany, and Denmark, among other places. However, it is impossible to judge from this map at what time of year users in these regions are using this emoji. In order to understand both where and when certain topics might have been discussed on Twitter, analysis of the temporal and spatial facets of the data needs to be conducted simultaneously.

3.4.4 Spatial-Temporal Typicality

Although the analysis of temporal and spatial typicality per emoji helped to identify trends in emoji usage, they each only illustrate half of the full picture. In order to gain an understanding of how emojis are used over both space and time, the temporal and spatial facets of the data needed to be analyzed in tandem. To achieve this goal, an additional spatial-temporal typicality workflow was conducted in which the spatial typicality of selected emojis was visualized over time. The result for each emoji is eleven

maps displaying spatial typicality per month which can either be viewed as a matrix (see Appendices D.2 - D.11), or as an animation (see Appendix E).

Of the sixteen emojis that were determined to be topically consistent, nine were selected for spatial-temporal analysis. A summary of which emojis were selected for the analysis is shown in Appendix D.1. A 100 by 100 kilometer grid was once again implemented to create spatial subsets of the data, however the subsets in this case were further specified to only include posts located within a grid cell that were published within a specified month. Complete typicality maps of the entire study area were therefore generated for each month in the dataset. When these maps are viewed in tandem, they allow for visual analysis of changing trends in emoji usage over both space and time. It is the author's opinion that the results are best viewed as animations because the dynamic visualization facilitates comparison of the same region over time. However, for static visualizations, a matrix of the individual frames (one for each month) can also provide a means for comparing across time and space following the methods used by Koylu (2019). The map multiples can be found in Appendices D.2 - D.11.

As discussed in McKittrick et al. (2022), a common method used in GIS-social media analysis is the comparison of results from a social media dataset with evidence from an alternative confirmatory source. This comparison demonstrated the validity of the findings by determining the degree to which the results from the social media data emulate the patterns observed in real-world events related to the given topic. An effort was therefore made to ground-truth trends found in the spatial-temporal analysis using evidence from real-world events. However, due to time constraints and the labor-intensive nature of the comparison, this additional investigation could only be conducted for a subset of the emojis selected for the emoji-specific analysis. This workflow involved identifying typicality hot-spots on the generated maps and researching corresponding events occurring at that time and location as well as researching large events related to each topic and searching for corresponding trends in the visualizations. The results of this analysis will be interpreted on an emoji-by-emoji basis in the following section.

3.4.5 Emoji-Specific Interpretation

The results of the topical consistency investigation indicate that not all emojis relate to specific topics. Indeed, many of the most commonly used emojis are popular because they convey broad emotions about a topic and do not represent a specific topic themselves. However, some emojis that represent concrete objects and ideas were found to correspond consistently with related topics.

In order to verify the results of the topical consistency workflow in some way, spatial and temporal trends in the use of thematically consistent emojis were compared to occurrences of related events in the real world. This form of ground-truthing revealed trends that support the use of emojis as proxies for certain topics on Twitter. Several examples of trends and their coinciding, real-world events are listed below. It should be noted that this workflow was not equally feasible for all emojis and was only performed

for emojis that were both topically consistent and displayed significant spatial, temporal, or spatial-temporal trends.

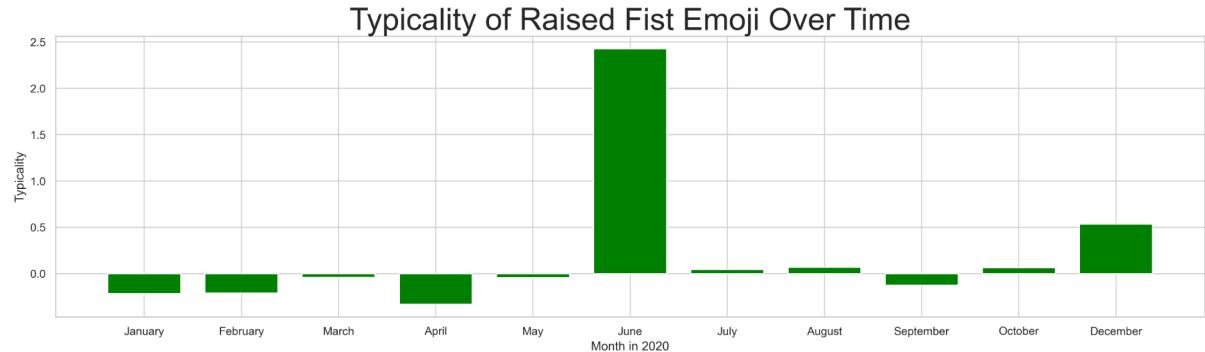
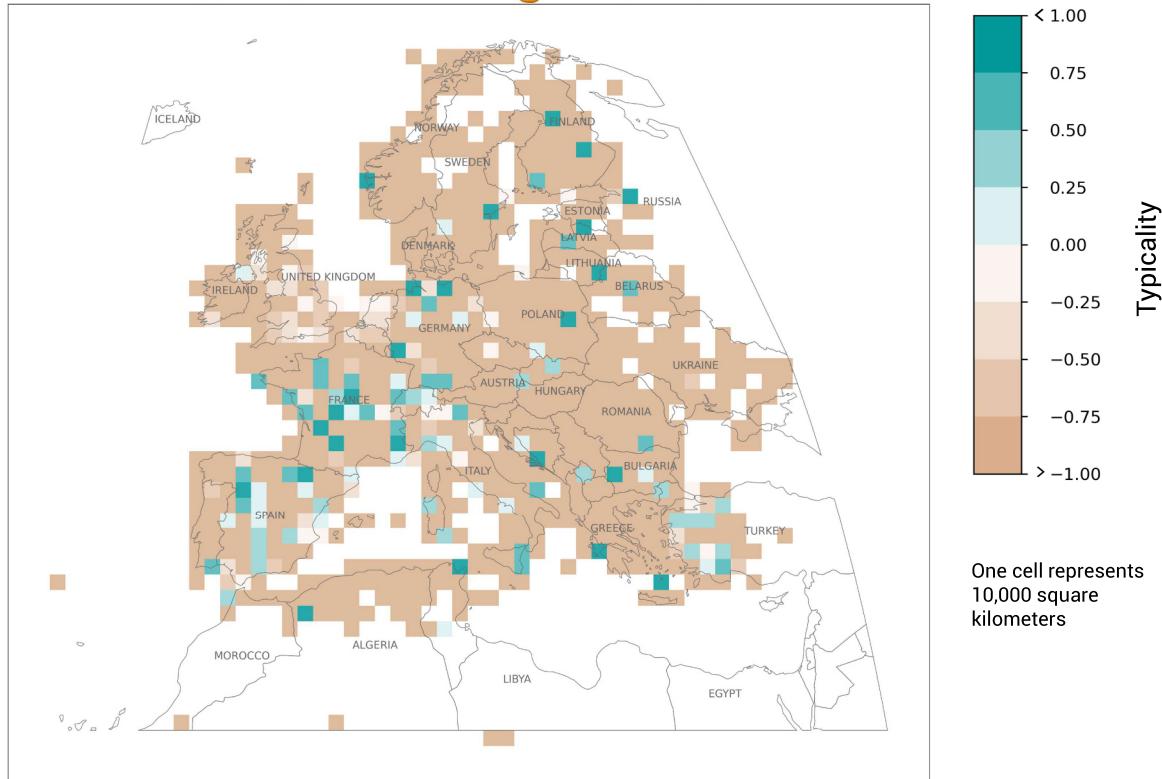


Figure 22: Typicality of the raised fist emoji (✊) over time.

The raised fist emoji was found to have a topical consistency of 90.94% concerning the Black Lives Matter (BLM) movement. Based on the results of the temporal typicality, it is clear that general increase in typicality of raised fist emoji occurs in June 2020 (see Figure 22). This phenomenon coincides with the international BLM movement gaining momentum online after the infamous murder of George Floyd by police officers in the United States in late May 2020. More specifically, in May of 2020, the results of the spatial-temporal analysis show the raised fist emoji being typical in the area surrounding Paris, France (see Figure 23). This phenomenon corresponds accurately with the discussion and organization of a demonstration on June 2 that took place despite a ban on gatherings of more than 10 people that existed at the time.

Typicality of Raised Fist 🤝 Emoji (May)

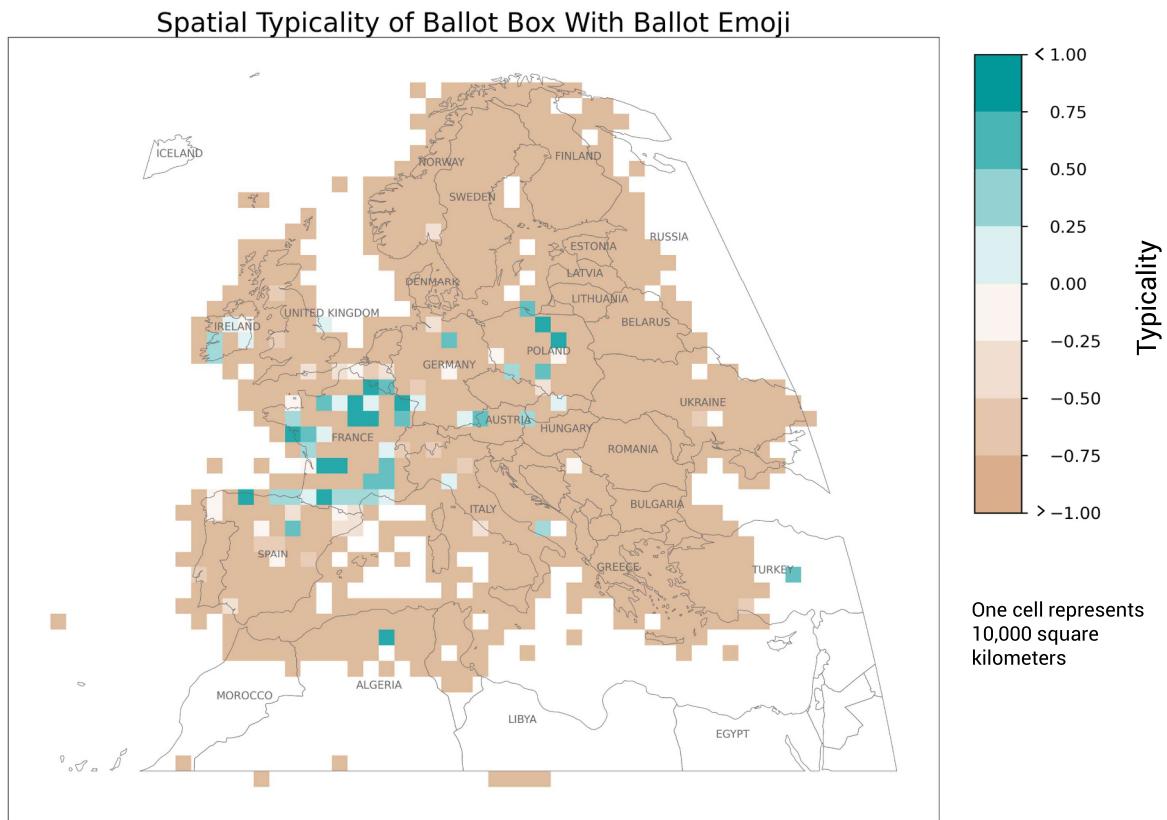


Data Sources: Natural Earth, Twitter database of the Technical University of Dresden Institute of Cartography

Author: Samantha Levi

Figure 23: Typicality map of the raised fist emoji (🤝).

The ballot box with ballot emoji (🗳️) also displayed a striking trend during spatial analysis (see Figure 24). The ballot box emoji demonstrated a topical consistency of 72.1% concerning elections and was typically found in France and Poland, two countries which held elections during the year 2020. In Poland, the presidential election was scheduled to take place in May but was postponed due to the pandemic. The first round of voting took place in late June, but since no candidate received a majority of the vote, a second round was held in July. In France, both senate and several municipal elections took place in the same year.



Data Sources: Natural Earth, Twitter database of the Technical University of Dresden Institute of Cartography

Author: Samantha Levi

Figure 24: Typicality map of the ballot box with ballot emoji (🗳).

The breastfeeding emoji (🤱) demonstrated a topical consistency of 88.8% concerning breastfeeding and babies. This emoji experienced a spike in typicality during the months of June, July, and August, as shown in Figure 25. This time frame corresponds with World Breastfeeding Week and similar campaigns that took place in July of 2020 to raise awareness and support for breastfeeding (Moukarzel et al., 2021).

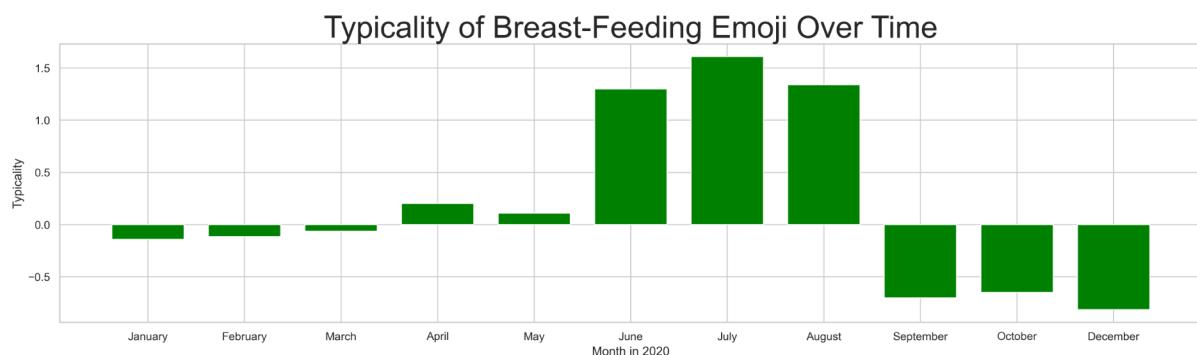


Figure 25: Typicality over time of the breastfeeding emoji (🤱).

The face with medical mask emoji (😷) has a topical consistency of 92.9% concerning COVID-19 and displayed a dramatic increase in typicality for the month of March and

remained typical throughout the months of April and May as shown in Figure 26. This trend coincides with the beginning of the COVID-19 pandemic in March 2020 and the subsequent establishment of mask mandates and other restrictions in many European countries. Indeed, the effect of the COVID-19 pandemic can be seen in the coinciding rise in typicality of the folded hands emoji (🙏), the 8 o'clock emoji (⌚), hospital emoji (🏥), and the woman health worker emoji (👷), which all have increases in typicality during March and April. The syringe emoji (💉) demonstrated an increased typicality at the end of the year in October and December due to a rise in discussions about vaccines.

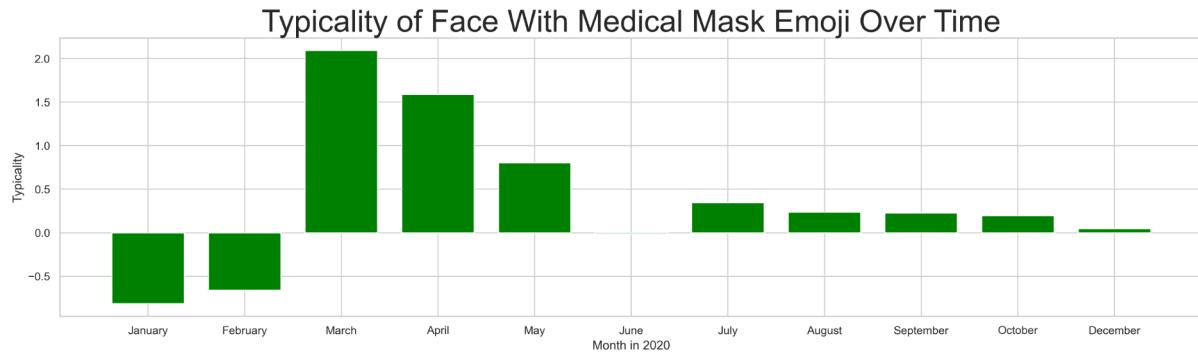


Figure 26: Typicality over time of the face with medical mask emoji (😷).

The rainbow emoji (🌈) has a topical consistency of 71.7% with regard to LGBTQIA+ rights and was found to be typical for the months of April, May, and June. This can be partially explained by the fact that June is international pride month, during which topics related to the LGBTQIA+ community are discussed. The rise in typicality during April and May may be due to the frequency of weather patterns that produce rainbows during these spring months.

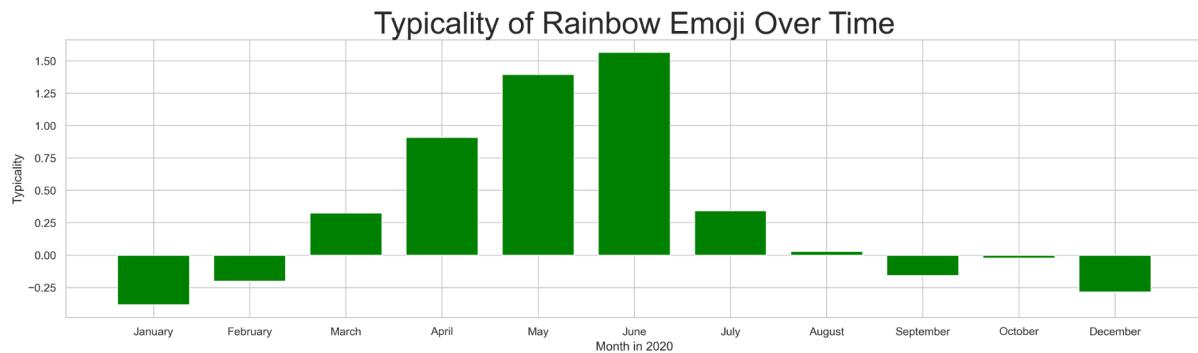


Figure 27: Typicality over time of the rainbow emoji (🌈).

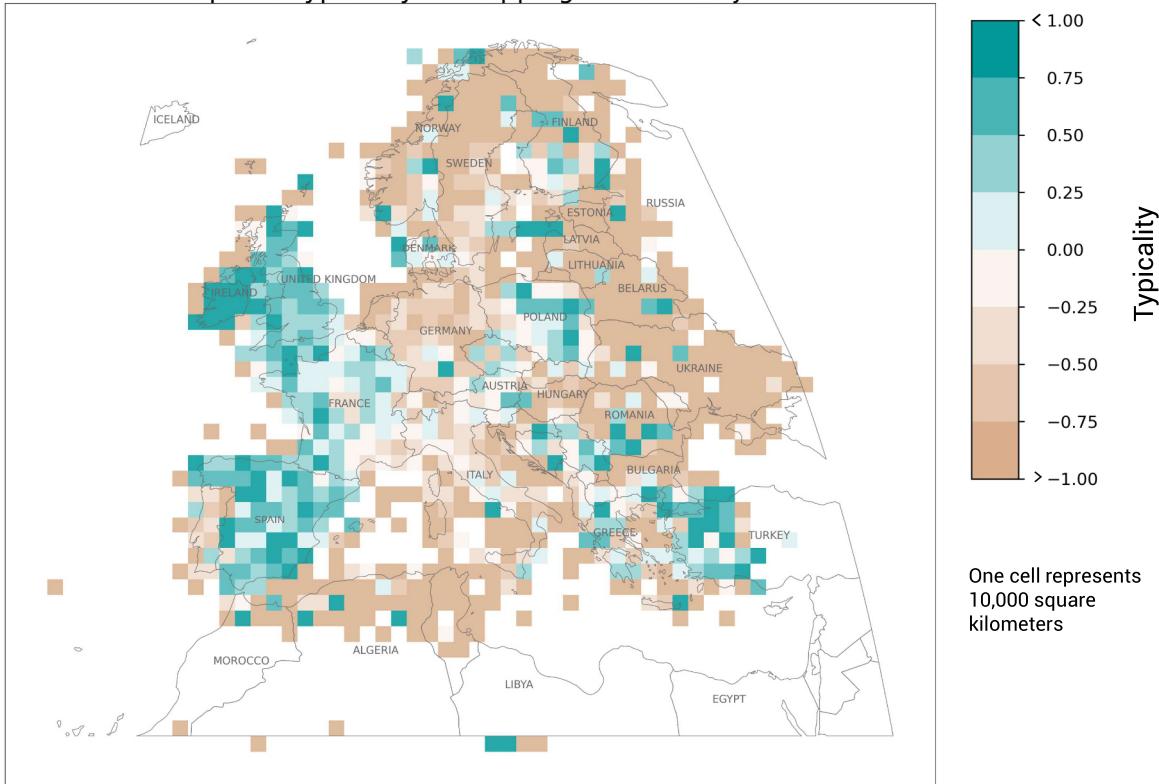
It should be noted that the list of topics found to be prevalent in this study is by no means exhaustive. It is entirely likely that additional topics were discussed extensively on Twitter, perhaps even more so than some of the topics mentioned here. This list does not necessarily represent the most frequently discussed topics, but rather topics whose discussion over time and space could be traced with the corresponding use

of topically-related emojis. The topics fall into three general categories: the COVID-19 pandemic and related themes, political discussions, and leisure activities.

The prominence of the COVID-19 pandemic as a discussion topic on Twitter was already prevalent in the early stages of exploratory analysis; #covid19, #coronavirus, and #covid_19 were some of the most commonly used hashtags in the dataset as shown in Figure 7. During the temporal typicality analysis, the typicality of several emojis spiked in correlation with the start of the pandemic in March 2020. Many of these emojis were confirmed to be topically consistent with respect to the pandemic, such as the face with medical mask (😷), microbe (🦠), hospital (🏥), and woman health worker (💉) emojis.

In addition to emojis that directly address the coronavirus, some emojis were topically consistent for other topics peripheral to the disease itself. For example, the clapping hands emoji (🙌) was topically consistent specifically to essential worker appreciation during months with the strictest lockdown regulations. Upon further investigation, the eight o'clock emoji (⏸), which was found to typically occur in Spain, was most often used to refer to the campaign of clapping for healthcare workers at 8 p.m. each evening. Similar applause campaigns took place in Spain, France, Ireland, and the United Kingdom, which is reflected in the typicality map of the clapping hands emoji in Figure 28.

Spatial Typicality of Clapping Hands Emoji



Data Sources: Natural Earth, Twitter database of the Technical University of Dresden Institute of Cartography

Author: Samantha Levi

Figure 28: Typicality map of the clapping hands emoji (👏).

Another prominent theme in the dataset was the discussion of political and human rights issues. The spatial typicality of the ballot box with ballot emoji (🗳️) correlates strongly with elections in France and Poland. The temporal typicality of the rainbow emoji (🌈) correlates with international LGBTQIA+ pride month, and the spike in spatial and temporal typicality of the raised fist emoji (✊) coincides with the international BLM movement.

Topics related to leisure and lifestyle activities were also successfully traced over space and time, including golf (through the use of the golfing man (🏌️) emoji), electronic music (through the use of the bomb (💣) emoji, which was found to be topically consistent for electronic music 72.3% of the time), and breastfeeding (through the use of the breastfeeding (🤱) emoji). The golfing man emoji was found to be typical in Ireland, the bomb emoji was typical in Italy (where it was used to refer to electronic music and specific DJs), and the breastfeeding emoji found to typically occur in the United Kingdom.

4 Results and Discussion

The overarching goal of this thesis, to determine whether emojis can be used to identify relevant topics and their spatial-temporal evolution in a non-topic-specific dataset, has been achieved. Based on the results of this research, it was determined that emojis can be used to identify relevant topics and their evolution within the given dataset as long as the emojis used are first found to be topically consistent. A methodology was presented for determining topical consistency using co-occurring hashtags to contextualize each emoji.

A major advantage of using emojis for spatial-temporal analysis of geo-social media is the relative simplicity of emojis compared with text. When compared with similar work that uses text-based analysis, this study was able to identify meaningful trends in the dataset with comparatively little computational effort. For example, Kruspe et al. (2020) implemented trained neural networks for sentiment analysis on a thematically-unfiltered dataset in order to determine overall sentiment of tweets across time and space during the COVID-19 pandemic. In comparison to the training and comparison of neural networks, this study's methodology is comparatively less computationally intense, due both to the lack of diversity of emojis the relatively straightforward computation of the typicality calculation.

Although the workflow was ultimately successful in identifying spatial-temporal changes in emoji use and their thematic connections, some limitations to the methodology should be noted. The chief limitation of the proposed method is that it is only applicable for Twitter data which contains both hashtags and emojis. This requirement significantly reduces the amount of data that can possibly be gathered for analysis, since many tweets lack one or both of these attributes.

The dissimilar rendering of emojis on different operating systems may also pose restrictions on the universal interpretations of emoji meanings. While the use of hashtags in combination with emojis helps to account for the possibility of semantic differences, this condition should still be considered when interpreting the results of this analysis.

Cultural differences in the interpretation of emojis may, to a lesser extent, also have an effect on emoji usage that is not analyzed in this study. However, the degree to which cultural differences affect the use and interpretation of emojis is not consistent and is certainly still less than the cultural differences in language encountered during text-based analyses. In a comparison of emoji usage between eastern and western countries, Guntuku, Li, Tay, and Ungar (2019) found that, while some variation in emoji interpretation exists across cultures, evidence implies a significant degree of universality of emoji interpretation across cultures.

One consideration that should be noted concerning the practical implementation of results is the relationship between privacy and functional use of data. User privacy must be considered in combination with the eventual application of collected data (Bender, Stodden, & Nissenbaum, 2014; Malhotra, Kim, & Agarwal, 2004). Although many

potential applications for the presented methodology exist, they are not all equal with regard to user benefits and are therefore also not all equal with regard to the level of user privacy that should be ensured. Because the purpose of this study was purely academic and intended to establish the validity of the proposed methodology, privacy-awareness via HLL formatting, cryptographic hashing, spatial aggregation, and coarse spatial visualization was deemed sufficient rather than total privacy preservation. Such a standard for privacy-awareness could also hold true for instances where significant user benefits exist, such as within the realm of environmental research (Andersson & Öhman, 2017; Dunkel et al., 2020). In instances where the applications of data can potentially save lives, as in the documentations of possible war crimes, even less protection of user privacy is sometimes accepted (Strick, 2022). However, in cases where users are unlikely to benefit from the applications of their data, additional privacy measures should be implemented.

The following subsections assess the degree to which the three main Research Objectives (RO) were achieved and answer the three corresponding Research Questions (RQ) posed in Section 1.2.

4.1 Research Objective 1

RO1: Develop a means for detecting change in emoji usage over time.

RQ1: Do significant changes in emoji use happen over time and space?

The variation in emoji use over both time and space was demonstrated throughout the analysis. Basic summary statistics of the absolute frequency of emojis on monthly and country-based subsets prove that these variations exist, while typicality analyses were able to uncover more specific spatial and temporal patterns within the dataset. Typicality calculations, being normalized for the size of the subset, were particularly suited to the temporal analysis given the temporal gaps in data collection since they allowed for the comparison of emoji usage across months with varying amounts of data.

Not all emojis demonstrated similar levels of variation. While some emojis were used fairly consistently over space and time, others varied greatly over one or both dimensions. The raw dataset was used to investigate the spatial and temporal facets of the data. HLL data format was used to narrow the scope of the analysis by generating lists of relevant countries and emojis with the most user days.

At this point it is essential to note that, while this dataset is as complete as possible given the available resources, inherent biases in the data exist. For example, in order for posts to be geotagged with a precise location, users must override Twitter's default settings and specifically agree to the use of their precise location. Therefore, the users represented in this dataset do not make up a random sample of the population; rather, they form a specific subset of users who are comfortable with these alternative

settings. By including only Tweets containing both emojis and hashtags, this dataset also represents only those users comfortable using both of these features. Individuals who are less comfortable with using emojis or hashtags, or who use social media less frequently for whatever reason will be unavoidably under-represented in this dataset. Specifically, users who enable the geotag functionality are more likely to be younger and higher income than the average user and be from more urbanized areas (Malik et al., 2021). Despite these unavoidable qualifications, the given Twitter dataset allows for a much broader scope of data than is possible through other methods at this time.

4.2 Research Objective 2

RO2: Determine whether significant changes in emoji usage over time and space correlate with significant topics.

RQ2: Do spatial and temporal changes in emoji usage have thematic connections?

In order to answer this research question, it first had to be determined whether emojis themselves have thematic connections. This study presents a methodology for the approximation of topical consistency using co-occurring hashtags and their occurrence numbers. Although a general threshold of 70% was established to denote topical consistency (based on the behavior of benchmark emojis representing concrete objects), the percent topical consistency of each emoji should be kept in mind when correlating emojis with their associated topics. Emojis with higher percentages of topical consistency can be used as more direct proxies for topics than emojis with lower percentages of topical consistency. The microbe emoji, for example, co-occurs with hashtags relating to the topic of COVID-19 approximately 98.7% of the time. The rainbow emoji, on the other hand, has a topical consistency of 71.7% with the subject of LGBTQIA+ rights. Between these two examples, the microbe emoji can be used as a much more reliable proxy for the discussion of COVID-19 than the rainbow emoji for the topic of LGBTQIA+ rights.

Ultimately it was found that many of the most frequently used emojis are not topically consistent. However, some topic-specific emojis were found that demonstrated topical consistency to varying degrees. The variations over time and space for these topically consistent emojis can potentially be used as a proxy for the discussion of specific topics over time. Several prominent topics and events found in the dataset include the COVID-19 pandemic, the Black Lives Matter movement, LGBTQIA+ rights, and various leisure activities. Such findings should not be used to generalize about the most popular topics for all Twitter users, but rather for the topics that were able to be detected through corresponding emoji usage amongst the subset of geotag users with their population biases (Malik et al., 2021).

4.3 Research Objective 3

R03: Visualize the results in a meaningful and comprehensive way.

RQ3: What are the most appropriate visualization methods to represent emoji usage over space and time?

For the representation of emoji use over time, both bar charts and line graphs were deemed suitable for analysis. These forms of data visualization are widely understandable and clearly illustrate trends in emoji usage over time. For both of these chart types, the data was listed in chronological order from left to right to facilitate interpretation. To display change in emoji use over space, maps were selected for visualization because of their "visual nature and high information density" (McKitrick et al., 2022). However, since the raw data format contains sensitive user information, several privacy concerns arose that had to be addressed. To avoid the visualization of precise user locations, for example, point mapping was not considered for final visualizations. Point mapping was only used for HLL data to display data that was already spatially aggregated. By dividing the study area into cells and assigning each tweet to a cell, spatial trends in emoji usage could be visualized without symbolizing individual points. The 100 by 100 kilometer grid that was created offers a very conservative level of data granularity to prevent the identification of individual users. Using typicality as a visualization metric as opposed to absolute frequency also mitigated the chance that individual users could be identified in cells containing very few data points.

The topical facet was not visualized on its own, but rather calculated independently and used to inform the selection of emojis for spatial-temporal analysis. To represent emoji usage over both space and time simultaneously, spatial typicality was visualized as a series of maps representing monthly spatial trends in emoji usage. These map series can be viewed either as static matrices or as dynamic animations for ease of comparison.

5 Conclusion

The results of this study support the use of emojis as indicators of spatial-temporal-thematic developments in geo-social media and illustrate the necessary considerations to be made when working with such data. Namely, the degree of topical consistency of each emoji should be taken into consideration when drawing comparisons between the use of emojis over time and space and the discussion of related topics on Twitter. Additionally, statistical measures that are easily skewed by hyperactive and non-human users, like absolute and relative frequency, are limited in their ability to derive meaningful insights from the data and should be avoided for the selection of relevant emojis. Metrics that are normalized across users, like typicality and user days, serve to minimize these influences. Awareness of user privacy should also be maintained wherever possible.

The 4 Facet structure of LBSN proposed by Dunkel et al. (2019) served as a suitable framework for this study. Using a combination of minimally processed "raw" data and the privacy-aware data format HyperLogLog (HLL), all four of the facets (social, spatial, temporal, and topical) were able to be analyzed, albeit not in tandem. The social facet was investigated through the use of HLL data, which helped to illustrate the use of emojis as well as provide insights as to which emojis are most used by hyperactive users. The spatial and temporal facets were explored using the geotagged coordinates and the publication date, respectively.

Typicality calculations were performed for spatial and temporal subsets of the raw dataset to gain insights as to the variation in emoji use over space and time. The topical facet was explored via emoji-hashtag combinations, and a methodology was proposed for the approximation of an emoji's topical consistency using the top 20 co-occurring hashtags with each emoji. For emojis which demonstrated significant topical consistency, further spatial-temporal typicality analysis was conducted. The resulting visualizations of this study highlight emojis that have significant spatial or temporal variations in use and which demonstrate approximate topical consistencies of over 70%. These visualizations have a minimalistic, straightforward design and can be viewed either as static matrices or as animations that facilitate comparison over time and space.

5.1 Future Work

The approach used in this study and the methodologies proposed for data analysis and exploration proved sufficient to identify relevant topics within the dataset. However, additional considerations could be implemented in the methodology to produce further insights. For example, some modifications that were made to the dataset to simplify calculations, such as the removal of skin tone modifiers on emojis and the removal of Regional Symbol Indicator letters representing flag emojis, could be eliminated in future research. A dataset without these qualifications could be used to determine whether similar emojis exhibit different spatial and temporal patterns depending on skin tone or whether relevant topics can be identified through the use of flag emojis.

The map visualizations of spatial and spatial-temporal typicality per emoji revealed a large number of cells with highly atypical emoji values. Given more time, this skew could be accounted for either with the normalization of the dataset or the implementation of alternative data classification methods, such as the head-tail break method used in (Dunkel et al., 2020).

Improvements could also be made to strengthen the statistical validity of the results presented in this study. While the results of the spatial-temporal typicality analysis allow for the non-arbitrary visual analysis of trends in emoji usage, further quantitative spatial-temporal analysis could be conducted using data clustering algorithms such as Density-based spatial clustering of applications with noise (DBSCAN). SaTScan, a software that uses space-time scan statistics to analyze spatial-temporal data, could also be used to detect statistically significant clusters of emoji usage.

References

- Andersson, E., & Öhman, J. (2017, 4). Young people's conversations about environmental and sustainability issues in social media. *Environmental Education Research*, 23, 465-485.
- Ayvaz, S., & Shiha, M. O. (2017). The effects of emoji in sentiment analysis. *International Journal of Computer and Electrical Engineering*, 9, 360-369.
- Bai, Q., Dan, Q., Mu, Z., & Yang, M. (2019, 10). A systematic review of emoji: Current research and future perspectives. *Frontiers in Psychology*, 10.
- Barbieri, F., Espinosa-Anke, L., & Saggion, H. (2016). Revealing patterns of twitter emoji usage in barcelona and madrid. *Artificial Intelligence Research and Development*, 239-244.
- Bender, S., Stodden, V., & Nissenbaum, H. (2014, 5). Privacy, big data and the public good: Frameworks for engagement..
- Broni, K. (2022, 7). *Global emoji use reaches new heights*. Retrieved from <https://blog.emojipedia.org/global-emoji-use-reaches-new-heights/>
- Chandra, R., & Krishna, A. (2021, 8). Covid-19 sentiment analysis via deep learning during the rise of novel cases. *PLOS ONE*, 16, e0255615.
- Chen, Y., Yuan, J., You, Q., & Luo, J. (2018, 10). Twitter sentiment analysis via bi-sense emoji embedding and attention-based lstm. In (p. 117-125). ACM.
- de Andrade, S. C., Restrepo-Estrada, C., Nunes, L. H., Rodriguez, C. A. M., Estrella, J. C., Delbem, A. C. B., & de Albuquerque, J. P. (2021, 1). A multicriteria optimization framework for the definition of the spatial granularity of urban social media analytics. *International Journal of Geographical Information Science*, 35, 43-62.
- Desfontaines, D., Lochbihler, A., & Basin, D. (2019, 4). Cardinality estimators do not preserve privacy. *Proceedings on Privacy Enhancing Technologies*, 2019, 26-46.
- Dunkel, A., Andrienko, G., Andrienko, N., Burghardt, D., Hauthal, E., & Purves, R. (2019, 4). A conceptual framework for studying collective reactions to events in location-based social media. *International Journal of Geographical Information Science*, 33, 780-804.
- Dunkel, A., Löchner, M., & Burghardt, D. (2020, 10). Privacy-aware visualization of volunteered geographic information (vgi) to analyze spatial activity: A benchmark implementation. *ISPRS International Journal of Geo-Information*, 9, 607. Retrieved from <https://www.mdpi.com/2220-9964/9/10/607>
- Feldman, L. B., Barach, E., Srinivasan, V., & Shaikh, S. (2021). Emojis and words work together in the service of communication.. Retrieved from http://workshop-proceedings.icwsm.org/pdf/2021_05.pdf
- Flajolet, P., Fusy, E., Gandouet, O., Meunier, F., Morales, C., & Welke, P. (2016). *Understanding the hyperloglog: a near-optimal cardinality estimation algorithm*.
- Gabarron, E., Dorronzoro, E., Rivera-Romero, O., & Wynn, R. (2019, 5). Diabetes on twitter: A sentiment analysis. *Journal of Diabetes Science and Technology*, 13, 439-444. Retrieved from <https://journals.sagepub.com/doi/pdf/10.1177/1932296818811679>
- Ghermandi, A., & Sinclair, M. (2019, 3). Passive crowdsourcing of social media in environmental research: A systematic map. *Global Environmental Change*, 55,

36-47.

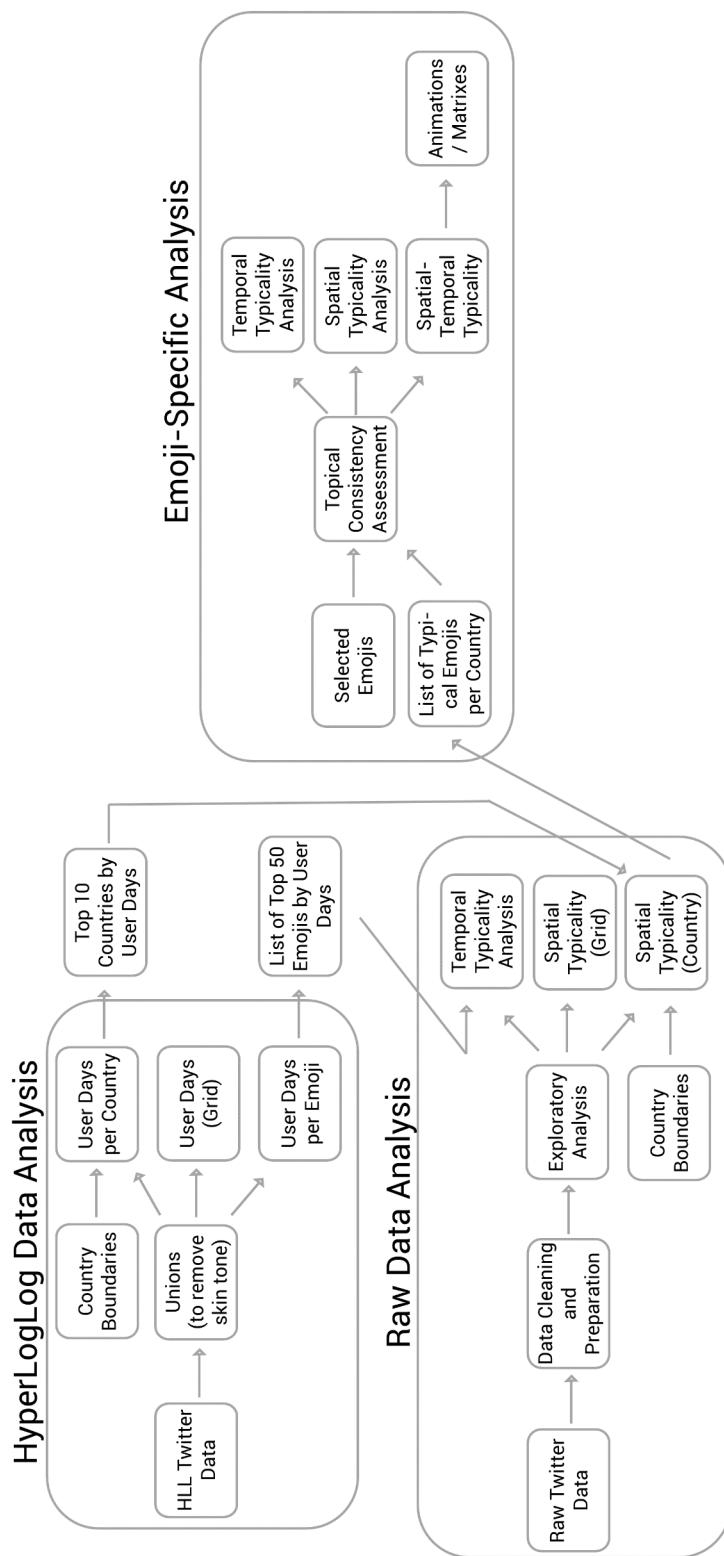
- Goodchild, M. F. (2007, 11). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69, 211-221.
- Granell, C., & Ostermann, F. O. (2016, 9). Beyond data collection: Objectives and methods of research using vgi and geo-social media for disaster management. *Computers, Environment and Urban Systems*, 59, 231-243.
- Guibon, G., Ochs, M., & Bellot, P. (2016, 6). From emojis to sentiment analysis. *WACAI*. Retrieved from <https://hal-amu.archives-ouvertes.fr/hal-01529708>
- Guntuku, S. C., Li, M., Tay, L., & Ungar, L. H. (2019). Studying cultural differences in emoji usage across the east and the west. In (p. 226-235). Retrieved from <https://ojs.aaai.org/index.php/ICWSM/article/view/3224>
- Hauthal, E., Burghardt, D., & Dunkel, A. (2019, 2). Analyzing and visualizing emotional reactions expressed by emojis in location-based social media. *ISPRS International Journal of Geo-Information*, 8, 113.
- Hauthal, E., Dunkel, A., & Burghardt, D. (2021, 6). Emojis as contextual indicants in location-based social media posts. *ISPRS International Journal of Geo-Information*, 10, 407.
- Imran, A. S., Daudpota, S. M., Kastrati, Z., & Batra, R. (2020). Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets. *IEEE Access*, 8, 181074-181090.
- International Organization for Standardization. (2013, 12). *Iso 19157:2013 geographic information – data quality*.
- Kejriwal, M., Wang, Q., Li, H., & Wang, L. (2021, 7). An empirical study of emoji usage on twitter in linguistic and national contexts. *Online Social Networks and Media*, 24, 100149. Retrieved from <https://doi.org/10.1016/j.osnem.2021.100149>
- Kim, K.-S., Kojima, I., & Ogawa, H. (2016, 9). Discovery of local topics by using latent spatio-temporal relationships in geo-social media. *International Journal of Geographical Information Science*, 30, 1899-1922.
- Kim, Y., ki Kim, C., Lee, D. K., woo Lee, H., & Andrada, R. I. T. (2019, 6). Quantifying nature-based tourism in protected areas in developing countries by using social big data. *Tourism Management*, 72, 249-256.
- Koylu, C. (2019, 4). Modeling and visualizing semantic and spatio-temporal evolution of topics in interpersonal communication on twitter. *International Journal of Geographical Information Science*, 33, 805-832.
- Kruspe, A., Häberle, M., Kuhn, I., & Zhu, X. X. (2020, 8). Cross-language sentiment analysis of european twitter messages duringthe covid-19 pandemic. Retrieved from <https://arxiv.org/pdf/2008.12172.pdf>
- Li, C., Sun, A., & Datta, A. (2012). Twevent: segment-based event detection from tweets. In *Proceedings of the 21st acm international conference on information and knowledge management* (pp. 155–164).
- Li, M., Chng, E., Chong, A. Y. L., & See, S. (2019, 9). An empirical analysis of emoji usage on twitter. *Industrial Management Data Systems*, 119, 1748-1763.
- Lin, T. J., & Chen, C. H. (2018, 8). A preliminary study of the form and status of passionate affection emoticons. *International Journal of Design*, 12, 75-90.

- Ljubešić, N., & Fišer, D. (2016). A global analysis of emoji usage. Association for Computational Logistics. Retrieved from <https://aclanthology.org/W16-2610.pdf>
- Löchner, M., Dunkel, A., & Burghardt, D. (2019, 10). Protecting privacy using hyperloglog to process data from location based social networks..
- Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004, 12). Internet users' information privacy concerns (uiopc): The construct, the scale, and a causal model. *Information Systems Research, 15*, 336-355.
- Malik, M., Lamba, H., Nakos, C., & Pfeffer, J. (2021). Population bias in geotagged tweets. In (p. 18-27). Retrieved from <https://ojs.aaai.org/index.php/ICWSM/article/view/14688>
- McKittrick, M. K., Schuurman, N., & Crooks, V. A. (2022). *Collecting, analyzing, and visualizing location-based social media data: review of methods in gis-social media analysis*. Springer Science and Business Media Deutschland GmbH.
- Moukarzel, S., Rehm, M., Caduff, A., del Fresno, M., Perez-Escamilla, R., & Daly, A. J. (2021, 3). Real-time twitter interactions during world breastfeeding week: A case study and social network analysis. *PLOS ONE, 16*, e0249302.
- Mukherjee, S. (2021). *Analyzing and visualizing location based social media data: The migration crisis in eu* . Retrieved from https://cartographymaster.eu/wp-content/theses/2021_Mukherjee_Thesis.pdf
- Openshaw, S. (1983). *The modifiable areal unit problem* (Vol. 38). GeoBooks.
- Prada, M., Rodrigues, D. L., Garrido, M. V., Lopes, D., Cavalheiro, B., & Gaspar, R. (2018, 10). Motives, frequency and attitudes toward emoji and emoticon use. *Telematics and Informatics, 35*, 1925-1934.
- See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., ... Rutzinger, M. (2016, 4). Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information, 5*, 55.
- Sloan, L., & Morgan, J. (2015, 11). Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. *PLOS ONE, 10*, e0142209.
- Strick, B. (2022, 5). *Eyes on russia: Documenting conflict and disinformation in the kremlin's war on ukraine*. Retrieved from <https://www.info-res.org/post/eyes-on-russia-documenting-conflict-and-disinformation-in-the-kremlin-s-war-on-ukraine>
- Twitter, Inc. (2020, 3). *Developer agreement and policy*. Retrieved from <https://developer.twitter.com/en/developer-terms/agreement-and-policy>
- Valley, R. W. L., Usher, A., & Cook, A. (2017, 10). Detection of behavior patterns of interest using big data which have spatial and temporal attributes. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-4/W2*, 31-35.
- Wiesław, W. (2016). Sentiment analysis of twitter data using emoticons and emoji ideograms. *The Central European Journal of Social Sciences and Humanities, 296*, 163-171.
- Wood, S. A., Guerry, A. D., Silver, J. M., & Lacayo, M. (2013, 12). Using social media to quantify nature-based tourism and recreation. *Scientific Reports, 3*, 2976.

Yu, Y. W., & Weber, G. M. (2020, 11). Balancing accuracy and privacy in federated queries of clinical data repositories: Algorithm development and validation. *Journal of Medical Internet Research*, 22, e18735.

Appendices

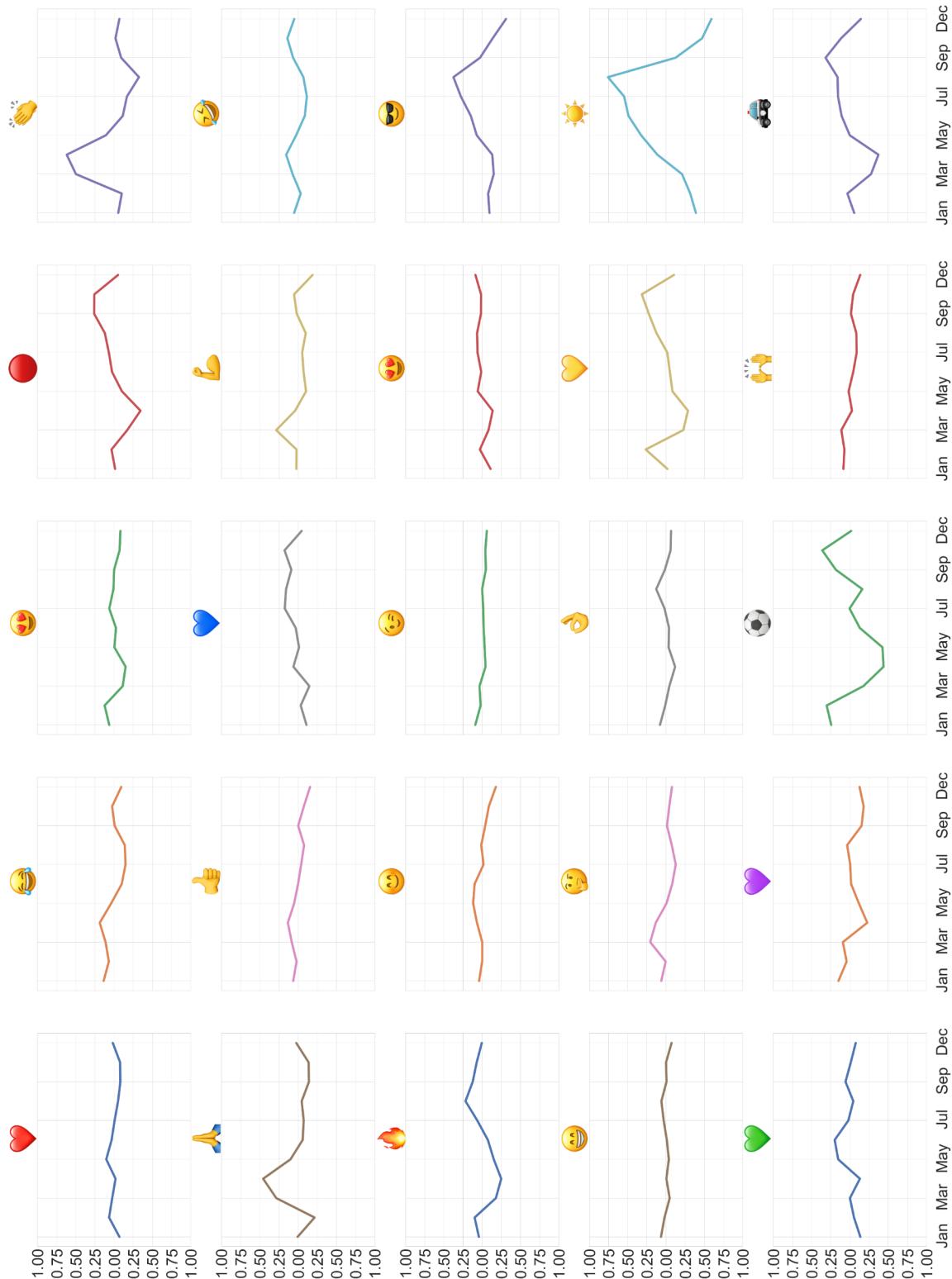
Appendix A Methods and Data Analysis



Appendix B Raw Data

B.1 Temporal Typicality for Top 50 Emojis by Absolute Frequency

Rank 1 - 25

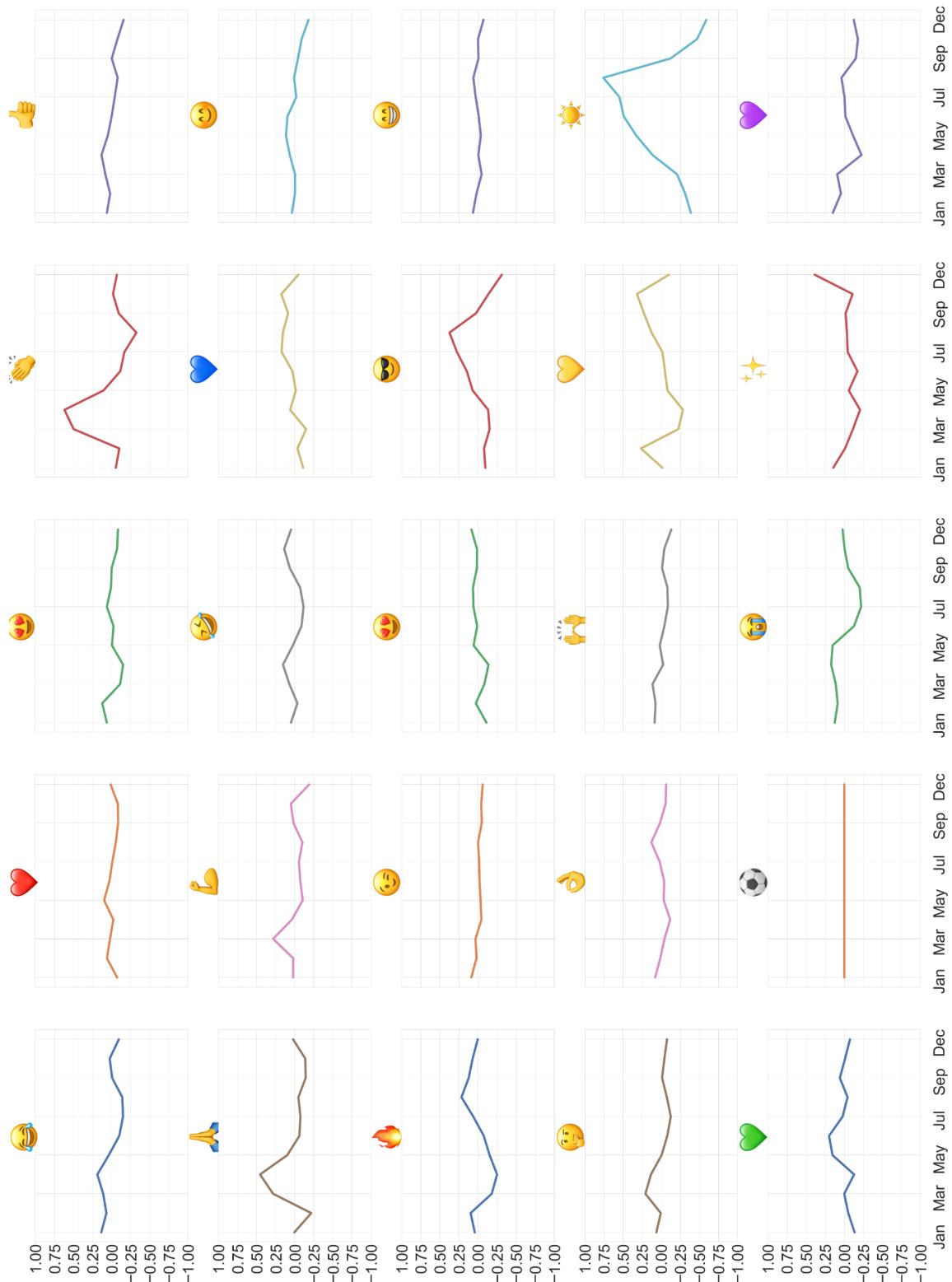


Rank 26 - 50

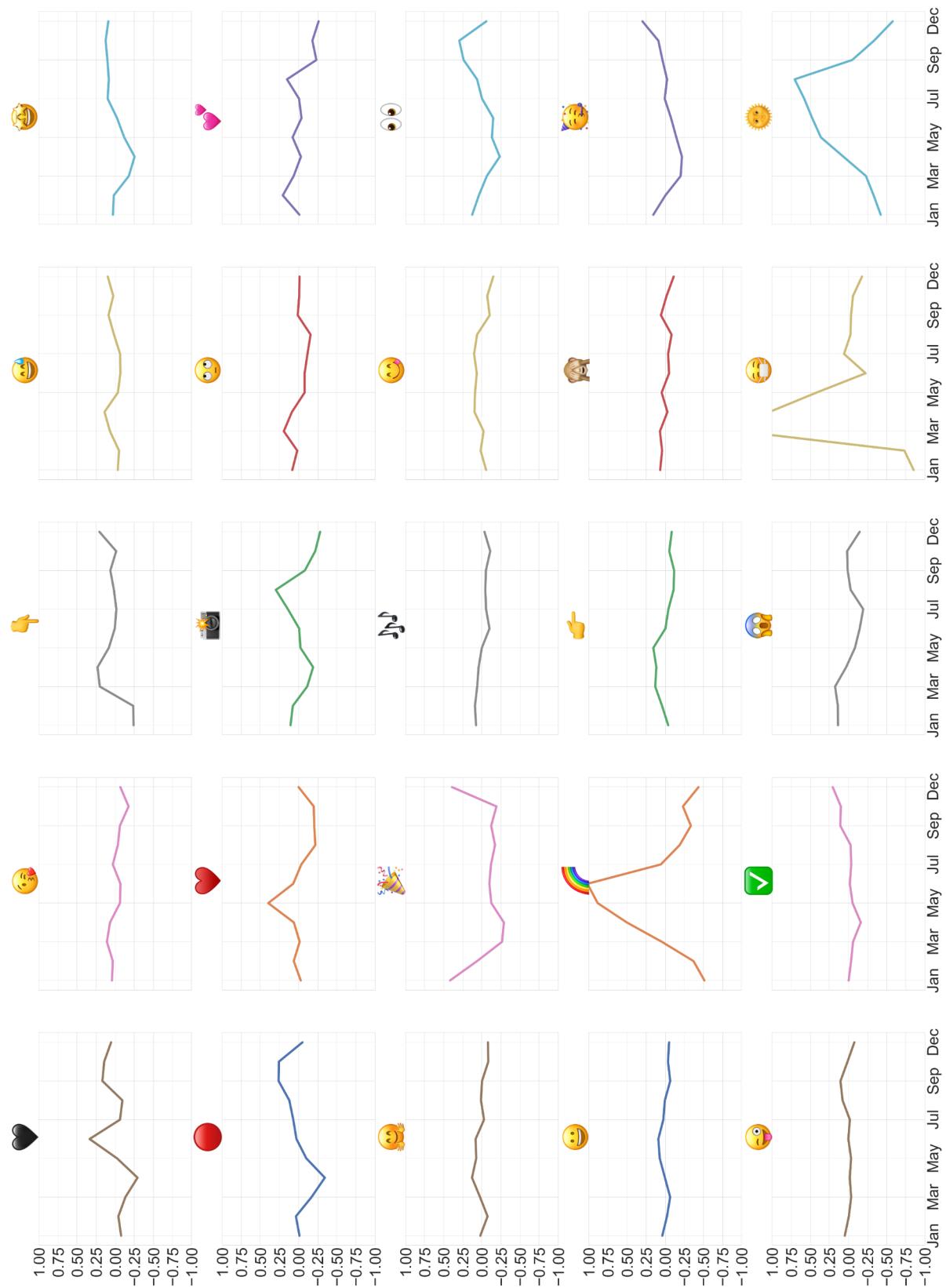


B.2 Temporal Typicality for Top 50 Emojis by User Days

Rank 1 - 25



Rank 26 - 50



Appendix C HLL Code

C.1 Cardinality Function

```
def hll_from_byte(hll_set:str):
    #ReturnHLLset from binary representation
    hex_string=hll_set[2:]
    returnHLL.from_bytes(
        NumberUtil.from_hex(
            hex_string,0,len(hex_string)))
def cardinality_from_hll(hll_set):
    #Turn binary hll into HLL set and return cardinality
    hll=hll_from_byte(hll_set)
    returnhll.cardinality()-1
```

C.2 Union Function

```
def union_hll(hll:HLL,hll2):
    """Union of two HLL sets. The first HLL set will be modified
       in-place."""
    hll.union(hll2)

def union_all_hll(
    hll_series:pd.Series,cardinality:bool=True)->pd.Series:
    """HLL Union and (optional) cardinality estimation from
       series of hll sets

    Args:
        hll_series: Indexed series (bins) of hll sets.
        cardinality: If True, returns cardinality (counts).
                      Otherwise,
                      the unioned hll set will be returned."""

    hll_set=None
    for hll_set_str in hll_series.values.tolist():
        if hll_set is None:
            # set first hll set
            hll_set=hll_from_byte(hll_set_str)
            continue
        hll_set2=hll_from_byte(hll_set_str)
        union_hll(hll_set,hll_set2)
    return hll_set.cardinality()
```


Appendix D Emoji-Specific Analysis

D.1 Summary of Analyzed Emojis

Emoji	Emoji Name	Topical Consistency	Topic	Typical Country	Typicality	Spatial-Temporal Analysis
🍺	Beer Mug	92.0%	Beer	-	-	-
🍷	Wine	75.4%	Wine	-	-	-
👉	Folded Hands	62.4%	Christianity	-	-	-
😷	Face with Medical Mask	92.2%	COVID-19	-	-	Yes
✊	Raised Fist	90.9%	Black Lives Matter	-	-	Yes
🌈	Rainbow	71.7%	LGBTQ+ Pride	-	-	Yes
💉	Syringe	71.4%	Vaccinations	-	-	-
🎄	Christmas Tree	87.9%	Christmas	-	-	-
🦠	Microbe	98.7%	COVID-19	-	-	Yes
👏	Clapping Hands	89.7%	COVID-19	-	-	-
🤱	Breast Feeding	88.8%	Childcare	United Kingdom	1.83	Yes
🧁	Cupcake	55.5%	Cupcakes	United Kingdom	1.14	-
🏌️	Golfing Man	86.9%	Golf	United Kingdom	1.13	Yes
⌚	8 O'Clock	78.0%	Essential Worker Appreciation	Spain	3.21	-
👩‍⚕️	Woman Health Worker	95.6%	COVID-19	Spain	2.66	Yes
🏥	Hospital	88.2%	COVID-19	Spain	2.18	-
🗳️	Ballot Box with Ballot	72.0%	Elections	France	3.25	Yes
💬	Speech Balloon	27.7%	-	France	3.19	-
↗️	Right Arrow Curving Down	34.7%	-	France	1.62	-
🤮	Face Vomiting	24.2%	-	Germany	1.42	-
👉	Index Pointing Up	33.6%	-	Germany	1.32	-
🤓	Nerd Face	55.6%	-	Germany	0.87	-
💣	Bomb	72.3%	Electronic Music	Italy	1.48	Yes
🤡	Clown Face	18.2%	-	Italy	1.07	-
젬	Gem Stone	56.5%	Jewelry	Italy	0.8	-
👄	Mouth	2.6%	-	Turkey	3.24	-
🔞	No One Under Eighteen	35.2%	-	Turkey	3.07	-
🎀	Ribbon	55.0%	Sex Work	Turkey	2.97	-
🚁	Helicopter	82.2%	Emergency Alert System	Netherlands	7.37	-
🚒	Fire Engine	69.2%	Emergency Alert System	Netherlands	0.22	-
🖤	Black Heart	24.7%	-	Belgium	0.63	-
🤔	Thinking Face	54.2%	COVID-19	Belgium	0.32	-
😎	Smiling Face with Sunglasses	22.9%	-	Belgium	0.12	-
❤️	Blue Heart	33.7%	-	Switzerland	0.11	-
😍	Smiling Face with Heart Eyes	13.3%	-	Austria	0.11	-

D.2 Ballot Box with Ballot Spatial-Temporal Matrix



D.3 Breastfeeding Spatial-Temporal Matrix



D.4 Clapping Hands Spatial-Temporal Matrix



D.5 Golf Spatial-Temporal Matrix



D.6 Hospital Spatial-Temporal Matrix



D.7 Mask Spatial-Temporal Matrix



D.8 Microbe Spatial-Temporal Matrix



D.9 Rainbow Spatial-Temporal Matrix



D.10 Raised Fist Spatial-Temporal Matrix



D.11 Woman Health Worker Spatial-Temporal Matrix



Appendix E Supplementary Materials

Additional materials, such as the Jupyter Notebooks containing the code used for analysis and visualization, data used for analysis, additional results, and a PDF version of this thesis can be accessed via the following public GitHub repository:

https://github.com/samsmop/EmojisAsIndicators_SupplementaryMaterials.