Cartography M.Sc.

Master thesis

Analyzing and Visualizing Location Based Social Media Data

The Migration Crisis in EU

Sagnik Mukherjee

Technical University of Munich TUM

TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology

TECHNISCHE UNIVERSITÄT DRESDEN

ITC UNIVERSITY OF TWENTE.

2021

Analyzing and Visualizing Location Based Social Media Data

The Migration Crisis in EU

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:	Sagnik Mukherjee
Study course:	Cartography M.Sc.
Supervisor:	Dr.-Ing. Eva Hauthal (TUD)
Reviewer:	Francisco Porras Bernárdez MSc. (TUW)

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

Date of submission: 10.09.2021

Statement of Authorship

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

Analyzing and Visualizing Location Based Social Media Data

I have fully referenced the ideas and work of others, whether published or unpublished.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

Dresden, 10.09.2021 Sagnik Mukherjee

Table of Contents

List of Figures

List of Tables

List of Equations

List of Codes

Acknowledgements

This master’s thesis would not have been possible without the priceless support from many different people. First and foremost, my gratitude goes to my first supervisor Dr.-Ing. Eva Hauthal. Not only would this thesis be impossible without her research, but also without her unparalleled support, her experience and knowledge and her confidence in my abilities. My gratitude also extends to Prof. Dipl.-Phys. Dr.-Ing. habil. Dirk Burghardt, who was at once interested to support my desire to have my own thesis topic. His guidance and wisdom have been motivating for me during this entire semester. Many thanks to Dr.-Ing. Alexander Dunkel who has always been available for questions and has been instrumental for setting up the resources required for this thesis. Of course, I must mention Juliane Cron M.Sc. for her constant advice and suggestions about everything under the sun throughout this entire two years of the Cartography program.

Living so far away from home has been made possible to the wonderful people from my Cartography intake. No matter what the situation was, I could find words of encouragement, help for assignments and friends for relaxing trips in them. Life in Europe would have been quite different without them all.

My special gratitude to my two dear childhood friends, Sayantan Ghosh and Srijan Uzir for being the best friends anyone could get. The conversations, comments, jokes and criticisms have and will continue to shape my life for days to come

Last but not the least life as I know it would be impossible without the ever-present love and support from my mother, father and aunt. Thank you for believing in me throughout my life.

Abstract

Over the past decade the use of social media has proliferated to the extent of enabling it to affect various activities of daily lives. One of the results of the extensive use of social media has been a new source of information for research into numerous aspects of social media data. Never in history did researchers have access to a rich and dynamic source of information of political and social behavior.

This thesis aims to add to the corpus of existing work while simultaneously applying new and developing techniques to define and understand online communicative practices on social media. The techniques include preserving user privacy, using hashtags for opinion analysis, using a framework of categorization developed to objectively capture events of social media, and using typicality as an indicator of said events. The case study chosen for applying them is the migration crisis in EU and offers unique challenges to applying these techniques to aid in their interpretation. The dataset gathered had tweets in six major European languages spanning from 2016 to 2021.

The result of this work shows that preserving the privacy of users leads to loss of contextual information and segregates the dataset such that it is unrecognizable from the original data. In other words, the privacy preserving HyperLogLog format could be better suited with real time data as compared to investigating events of the past, as was the author’s case. The dashboard designed to visualize some of the results works well at giving users an overview of the dataset but can be improved further.

A low number of tweets prevents the author from making conclusive remarks about using hashtags for opinion analysis. However, cases have been found where co-occurring hashtags can be used to broadly indicate positive or negative attitudes. This is very useful when certain hashtags of one language are widely used in different languages. It is also evident that most of the peaks found on the trendlines of tweets were related to political disturbances in the respective countries of tweet origin. Evidence of use of hashtags which were popularized by established institutions were also obtained for both pro-refugee and anti-refugee sentiments.

1. Introduction

1.1 Motivation and Problem Statement

Social media has, throughout the course of the last decade changed how people interact with each other. Gone are the days when communicating would mean making phone calls, messaging and written communication have replaced (Arnett 2015) . The trend was also reflected on the internet as social networking sites like Twitter allowed users to express themselves primarily through words.

The mode of such communication had few similarities to post cards or letters, the major forms of distant communication 50 years before. Rather new methods were developed for users to be able to express themselves in fewer words and be able to simulate non-verbal communication. One of these new methods are emojis, first being popular in Japan (Blagdon 2013) and then exploding onto the world. They are now a standard toolkit for written communication not only included in messaging services, but also for public posts on social media platforms.

Another ubiquitous tool allows for users to relate to different topics by means of keywords. Such keywords also serve to make posts relevant to events being discussed by the global internet community. Now being used on all major social media platforms, a true child of the internet and Twitter (Parker 2011), hashtags will play an instrumental role in filtering and analyzing tweets for this research.

Armed with these novel tools, users allowed themselves to react and discuss on events of personal and societal importance. The dynamics of tweets and retweets cross over from traditional spatial and temporal boundaries (Pruthi et al. 2015). Hence, making Twitter an important platform which sheds light on various events and the dynamics of public perception on such events.

The author plans to use Twitter to analyze public opinions and reactions to the EU migration crisis. The crisis itself has two very interesting features, which makes the topic unique for a case study:

- The crisis affects two very different sets of people; the ones migrating and the ones who are facing the migration.
- The crisis does not have, yet, a definite end point.

With the support and active cooperation of the Sächsicher Flüchtlingsrat (SFR), this work plans to improve our understanding of the public perception and attitude of Europe towards the influx of refugees and the subsequent crisis, which is an important issue all over the world.

1.2 Research Objectives and Questions

The unique characteristics of the migration crisis in EU gives us an opportunity to gauge how public opinion sways in geographic space and time as the fabric of society and community is challenged. For this purpose, tweets relevant to the migration crisis will be analyzed. Further, the objective would be to find possible causes for the patterns obtained from the data. This work would also include a suitable visualization for the analysis and variations of opinions from the results. This could be a static or an interactive visualization or even a combination of both, depending on the results obtained. Furthermore, a novel privacy-aware data format, the HyperLogLog format, will be tested on this dataset.

It must be noted that the work will chiefly focus on opinion analysis of the residents of Europe and not the refugees. The tweets extracted are in the major European languages and not in the languages that refugees would be expected to tweet in.

Considering the objectives briefly described above, the research questions can be summarized as follows

1. How does the HLL format perform for the EU Migration Crisis Twitter?
2. How can the results of the analyses be clearly and concisely visualized?
3. What conclusions can we draw from the analyses?

1.3 Innovations

The work is aiming to apply data-driven techniques on data including all of Europe from the year 2016 to 2021. Previous works on the topic of migration crisis in EU and using emojis for sentiment analysis (Yuita Arum Sari et al. 2014) have focused on smaller temporal windows. In addition to the temporal window, the use of six different European languages, to the author's best knowledge, is something that has not been previously attempted. The use of the privacy aware data format along with opinion analysis of the tweets is, according to the author, what makes the work innovative and unique.

1.4 Outline of the Thesis

Following the introduction, there will be four further parts with their own subdivisions in this work. The next part will be the Background, which goes over some past works relevant for this thesis. The Methodology section is going to describe the techniques employed to get the results. This will be followed by a Discussion of the results with regards to the objectives set out. Finally, the Conclusion and Outlook section will end with the final remarks of the author and discuss future possibilities of this work. The Appendix section has links for the supplementary materials.

2. Background

This section goes briefly over some of the past work that this thesis is going to be based on. The broad topics to be discussed here are the HyperLogLog format, facets of Location Based Social Network (LBSN), Typicality, Opinion analysis and past work on the migration crisis on social media.

2.1 HyperLogLog (HLL)

The HyperLogLog (Flajolet P. et al., 2007) is a solution for the count-distinct problem (Leskovec et al. 2020). The algorithm allows quantitative analysis of large datasets with an error rate of 2% (Flajolet P. et al., 2007). Coupled with geo-privacy preserving principles, the HLL formatted data allows for greater preservation of privacy of users compared with raw LBSN data.

The concern amongst users on social media for risks to privacy has increased in recent years. According to a 2014 poll by the Pew Research Center, 91% of the respondents "agree" or "strongly agree" to consumers losing control over how companies collect and use personal information from social media (Madden 2014) . In 2020 the European Union Agency for Fundamental Rights published a report in response to governments in the EU discussing the use of technologies to stem the spread of COVID-19. According to this report 41% do not want to share their personal data with companies, which is almost double the number compared to public bodies (European Union Agency for Fundamental Rights 2020) .

The benefit of privacy is only a side effect of this algorithm which is achieved through multiple steps to prepare the data for analysis. Publicly available Volunteered Geographic Information (VGI) data is processed into the HLL format (Löchner et al. 2019). This processing implements abstraction as used by cartographers to represent spatial data in various scale dependent resolutions (Burghardt et al. 2014). Furthermore using cryptographic hashing of data and increasing spatial granularity by using GeoHash (Dunkel et al. 2020), the format reduces risks by following geoprivacy-by-design recommendations (Kounadi et al. 2018). These techniques also mitigate the risks from intersection attacks (Desfontaines et al. 2018).

Along with reducing risks to breach of privacy of users, the HLL format also allows for further analysis of the data. Meaningful observations and conclusions can be drawn using metrics or overlays. The metrics used are post count, user count and post user days. Post user days is defined as the total number of days, across all users, that each user took one photograph within each site (Wood et al. 2013). In addition to the metrics, HLL format data permits set theory operations like union and intersection. Every HLL set can be combined, and the cardinality calculated to show the distinct number of the metric (post count, user count or post user days) being counted. Similarly, an intersection operation can also be performed for unearthing patterns of the metrics. This gives analysts some options for qualitative analysis (Dunkel et al. 2020).

2.2 Facets of LBSN

Facets are used to characterize reactions to events in LBSN (Dunkel et al. 2019). The authors define an individual reaction from a single actor to a single event as a reaction that cannot be further differentiated in a meaningful way. They define a reaction as a tuple as

$$r = (e, p^r, t^r, s^r, a^r)$$

Equation 1 Definition of a reaction in LBSN

where,

e stands for the event which is being reacted to.

p^r stands for actor who reacted. This is the social facet. The actor here can not only be an individual, but also institutions and organizations. This facet also considers the wider socio-political ecosystem that the actor belongs. LBSN does not directly capture the complex relation of the actor in their ecosystem, but these relations can be inferred from other data like the language, gender, age, etc.

t^r stands for the time of the reaction which is called the temporal facet. This facet contains not only the timestamp of the LBSN post, but also when the posts were modified or when the user joined the social media platform

s^r stands for the location of the reaction which is known as the spatial facet. Together with the temporal facet, the spatial facet forms the basis of characterizing a reaction and inferring the event.

a^r stands for the for the combination of attributes which define the reaction. This is known as the thematic/topical facet. The topical facet encompasses the situational reaction of the actor. It may include sentiments expressed (Hauthal 2015) or inferring the disposition of the actor from the title, comments or descriptions of the post (Zeng et al. 2016).

The facets not only provide a conceptual model for analyzing LBSN reactions, but also provide abstraction layers for added privacy protection by adjusting the granularity of the data (Löchner et al. 2018).

2.3 Typicality

Typicality is derived from the adjective typical which describes the "state of being that is typical." The formula was developed to combine two relative frequencies : one of a sub-dataset and the other of the total dataset of emojis or something similar. (Hauthal et al. 2021). The typicality t is described by

$$t = \frac{f_S - f_t}{f_t}$$

Equation 2 Definition of typicality

where, f_s is the relative frequency of the sub-dataset and f_t is the relative frequency of the total dataset. Both frequencies are calculated using

$$f = \frac{n}{N}$$

Equation 3 Definition of relative frequency

Where n stands for the absolute number of a specific emoji or hashtag and N stands for the absolute number of all emoji or hashtag.

Typicality has either a positive (typical) value or a negative (atypical) value. Compared to other tests for measuring normalization and relative differences, typicality is rather straightforward and easy to comprehend.

2.4 Opinion Analysis with Social Media Data

The work by Hauthal et al. (2019) will serve as a starting point for the thesis. It describes the approaches to opinion analysis with emojis and hashtags, which will be helpful. Additionally, the visualizations used would similarly be a starting point for exploring the visualization options to be used by the author. This is a long scale event which might be useful for this work as well (Hauthal et al., 2019). The case study, however, is completely different as this work will be focusing on the migration crisis in EU. Also, a primary investigation of the tweets to be used, shows a rather low number of emojis used for the posted tweets. Therefore, the data might not be suitable for a thorough sentiment analysis as illustrated by the cited work.

Effects of emojis on sentiment analysis of tweets are quite well described by Ayvaz and Shiha (2017). This paper could be a reference on the expected effects of emojis on tweets. However, the case study presented, and the methodology used provides only an expected outcome. For a thorough exploration for sentiments Kralj Novak et al. (2015) will be very helpful. Li et al. (2019) report that face with tears of joy and the heart emoji are the most abundantly used emojis on Twitter. They also assign sentiment scores to emojis and have found that positive emojis are used overwhelmingly more than negative ones. It would be interesting to see if these results hold true in the case of the migration crisis in EU as well.

2.5 EU Refugee Crisis and Social Media

Similar work on the migration crisis in EU and public perception has been presented by Inuwa-Dutse et al. (2020) . The paper discusses public perception of the migration crisis from a global dataset of tweets. The authors report a prevalence of negative sentiments expressed on Twitter in Europe and North America from unverified users and a generally positive or neutral sentiment expressed by verified users (verified users are defined as users having sufficient followers on their Twitter account. Öztürk and Ayvaz (2018) have worked on the sentiments expressed on Twitter for the Syrian refugee crisis in English and Turkish. They report finding a higher percentage of negative sentiments in English as compared to Turkish. Both the works discussed above are consistent in finding a higher percentage of negative sentiments about the migration crisis in English. This is something that would be expected in this thesis. However, it would be interesting to also see the opinions expressed in other European languages and compare them to that in English.

2.6 Visualizations

There are mainly two broad approaches for data visualization: static and interactive visualizations (Mahajan and Gokhale 2018). The dataset of this thesis required both kinds of visualizations to represent the findings. There was the need for visualizing the facets independently and in relation with each other. The facets independently were visualized using trend lines, word clouds and heatmaps as specified by Kumar et al. (2014). These visualizations, for example a static map, cannot encode all or even some of the facets along with the spatial facet.

To mitigate the problem of loss of contextual information interactive visualizations would be the solution. This is because of features like tooltips which allow additional information to be presented to the user (Heer et al. 2008). Such an interactive data visualization (Janvrin et al. 2014) also accommodates elaborate or abstracted views and filtering of data based on specific conditions for users (Yi et al. 2007). Interactive visualizations have also been shown to improve cognition for complex datasets (Dix and Ellis 1998) .

3. Methodology

3.1 Data Collection and Filtering

Since the aim was to collect geo-tagged tweets relevant to the migration crisis, hashtags were used to filter semantically relevant tweets from the non-relevant tweets. Being a pan-European phenomenon, it was chosen to broaden the dataset by including multiple languages. The choice of the language also depended on the data source, which in the author's case was not a live Twitter feed. Since the author intended to study Twitter data from the beginning of the crisis, or as close to the beginning as possible: the Tweet database of the Institute of Cartography was queried.

Initially, an attempt was made to include tweets in English, Spanish, Italian, Greek, German, Russian, French, Dutch and Turkish. The reasoning was to try to improve the number of tweets by also including languages where the migrants would be arriving (Turkish and Greek) but would not be major European languages. It must also be noted that recommendations of hashtags were also asked from SFR. These hashtags were in German and English and included a wide range of topics from deportations to Afghanistan (which became relevant recently) to general hashtags with respect to refugees and migration. Thus, an initial query was made using this following code

```
SELECT tweet_id, user_id, regexp_replace(tweet_text, E'[\\n\\r]+', ' ', 'g' ) AS tweet_text, tweet_lang, created, user_screen_name, user_lang, user_location, coord_type, place_lat, place_lng, native_lat, native_lng, source

FROM geotweet

WHERE (LOWER(tweet_text) LIKE '%abgeschob%'

OR LOWER(tweet_text) LIKE '%abschieb%'

OR LOWER(tweet_text) LIKE '%asiel%'

OR LOWER(tweet_text) LIKE '%asile%'
OR LOWER(tweet_text) LIKE '%asilo%'

OR LOWER(tweet_text) LIKE '%asyl%'

OR (LOWER(tweet_text) LIKE '%deport%' AND (LOWER(tweet_text) NOT LIKE '%deporte%' AND tweet_lang NOT LIKE 'es'))

OR LOWER(tweet_text) LIKE '%déport%'

OR LOWER(tweet_text) LIKE '%flücht%'

OR LOWER(tweet_text) LIKE '%göçmen%'

OR LOWER(tweet_text) LIKE '%iltica%'

OR LOWER(tweet_text) LIKE '%migración%'

OR LOWER(tweet_text) LIKE '%migrant%'

OR LOWER(tweet_text) LIKE '%migrati%'

OR LOWER(tweet_text) LIKE '%migrazione%'

OR LOWER(tweet_text) LIKE '%mültecil%'

OR LOWER(tweet_text) LIKE '%profugo%'

OR LOWER(tweet_text) LIKE '%refugee%'

OR LOWER(tweet_text) LIKE '%refugiad%'

OR LOWER(tweet_text) LIKE '%réfugié%'

OR LOWER(tweet_text) LIKE '%rifugiati%'

OR LOWER(tweet_text) LIKE '%sığınmacı%'

OR LOWER(tweet_text) LIKE '%sınır dışı%'

OR LOWER(tweet_text) LIKE '%vluchteling%'

OR LOWER(tweet_text) LIKE '%απελα%'

OR LOWER(tweet_text) LIKE '%απέλα%'

OR LOWER(tweet_text) LIKE '%ασύλο%'

OR LOWER(tweet_text) LIKE '%άσυλο%'

OR LOWER(tweet_text) LIKE '%μετανάστ%'

OR LOWER(tweet_text) LIKE '%πρόσφυγ%'
OR LOWER(tweet_text) LIKE '%бежен%'
OR LOWER(tweet_text) LIKE '%беженец%'
OR LOWER(tweet_text) LIKE '%депорт%'
OR LOWER(tweet_text) LIKE '%мигрант%'
OR LOWER(tweet_text) LIKE '%миграци%'
OR LOWER(tweet_text) LIKE '%убежищ%');
```

Code 1 Query for extracting relevant tweets

Before divulging into the results of this initial query it would be worthy to note that for the author’s intended purposes, there would be two kinds of filtering required on the data. As is clear from the query, exact hashtags or words within the tweet body are mostly not searched. Instead, words or hashtags similar to the initial list of hashtags are being searched. Hence there always remains a possibility that the search will return words that include our search term but semantically is unrelated to the issue of migration in the EU. This leads to the need for another step of filtering. The second kind of filtering would be a semantic filter. For example, searching for words or hashtags with the word *asylum* would return *asylum16,* which is unrelated to the migration crisis. The method for determining semantic suitability would be to use the hashtag explore (# explore) function on twitter to check the contents associated with the hashtag or word.

From this initial query, began another process of filtering using the two principles outlined above. Furthermore, the initial query also returned plenty of tweets which had other co-occurring semantically relevant hashtags. These hashtags were also searched within the returned tweets to estimate their quantitative occurrence. If these hashtags did occur many times (usually more than a 100 tweets) within the dataset, then they were counted and included in the list. This was also performed for the hashtags received from SFR. Also, a list of words was compiled to be specifically excluded from the query so as to not have any further (the hashtag e**asyl**ike would appear because of the German/Dutch hashtag **asyl**). A summary of the entire process of filtering is illustrated in Figure 4.

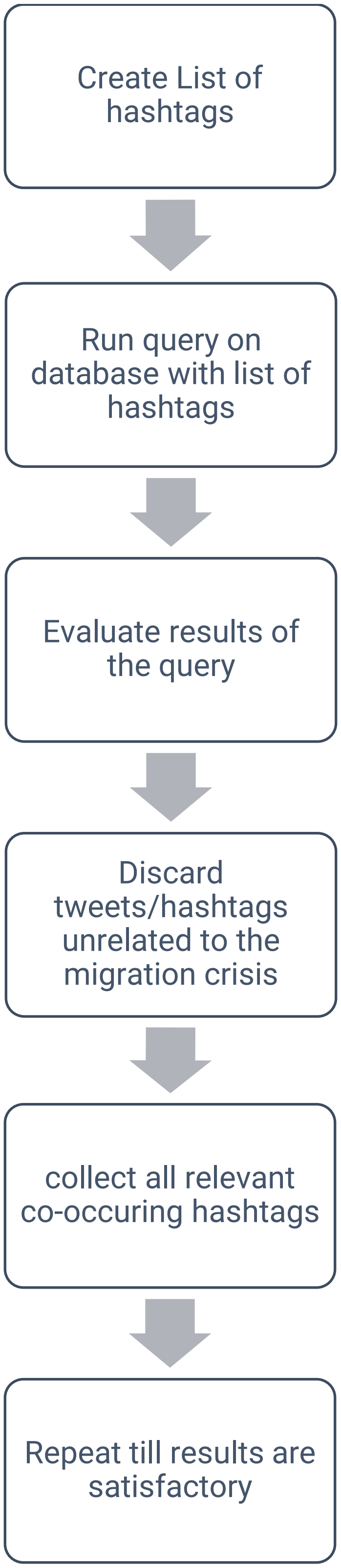

Create List of hashtags
Run query on database with list of hashtags
Evaluate results of the query
Discard tweets/hashtags unrelated to the migration crisis
collect all relevant co-occuring hashtags
Repeat till results are satisfactory

Figure 1 Summary of Data Collection and filtering

The table 1 shows the first 27 rows of the total 100 rows of hashtags and their counts. The filter terms in table 2 refer to the words that were excluded to keep the query as clean as possible.

Hashtag	Count
refugees	16530
migrant	14122
Migrants	12213
Refugeeswelcome	11807
Migration	8838
Immigration	7416
Refugee	4515
Refugiados	3741
Immigrazione	3035
Flüchtlinge	2991
Worldrefugeeday	2977
Withrefugees	2834
Asylum	2145
Refugeecrisis	2665
Migranten	2084
Immigrati	2015
Migrant	1871
Immigrants	1794
Réfugiés	1725
Refugeegr	1714
Vluchtelingen	1665
Migrantcrisis	1130
Rifugiati	1034
Migrationeu	1012
Asyl	934
Welcomerefugees	919
Refugeeweek	896
Asylumseekers	790

Table 1 List of hashtags prepared for second round of filtering

Filter terms
brasil
portugal
Basil
Deportivo
Easyl
Lunatic
deportivas
Lasilenciosacat
Almasyletras
Elazığsilerilstifa
Asilopercani
Asylum16
Asylum18
Asylum17
Mdtrefugiados
Fantasy
Emigrates
Arkhamasylum
Ahoravasylocascas
Easilocks
Letrasylatidos
Emigration
asilonido
Asilobelgaarv
Asilentvoice

Table 2 List of hashtags used to clean the query

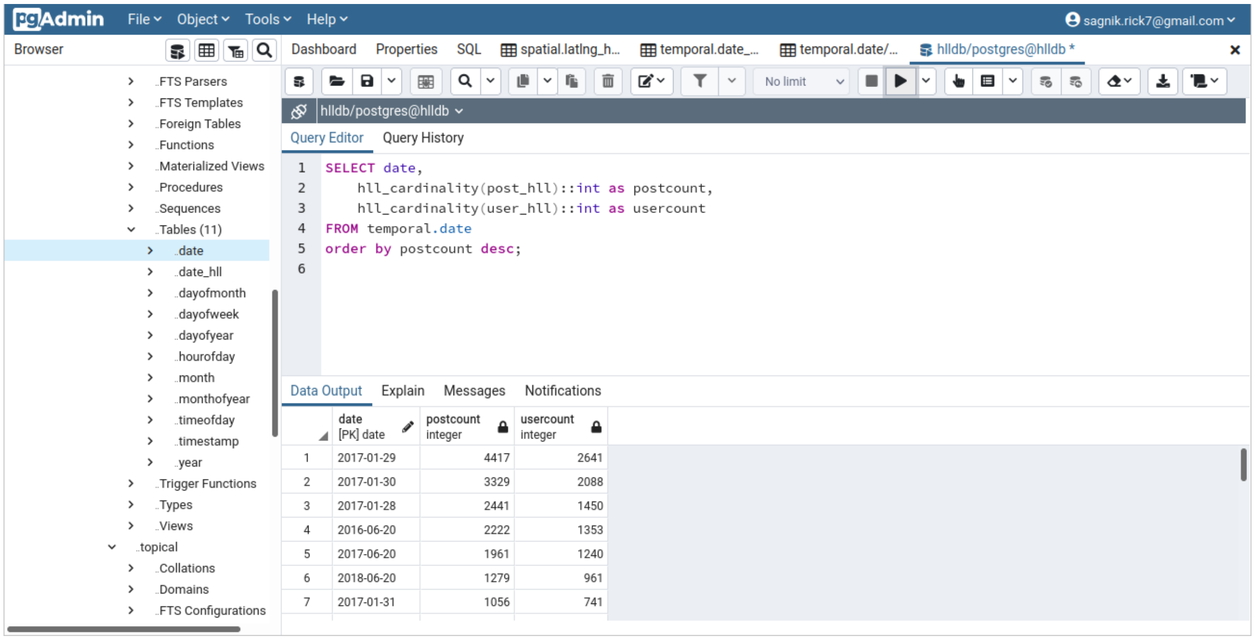
The results of this query were then stored in two separate servers on a PostgreSQL database: One for the Raw data format and the other for the HLL format to be analyzed and explored.

3.2 Data Exploration

As previously mentioned, data extracted from the HLL and Raw servers were explored separately. A brief overview of the techniques utilized is given in the following section.

3.3.1 HLL

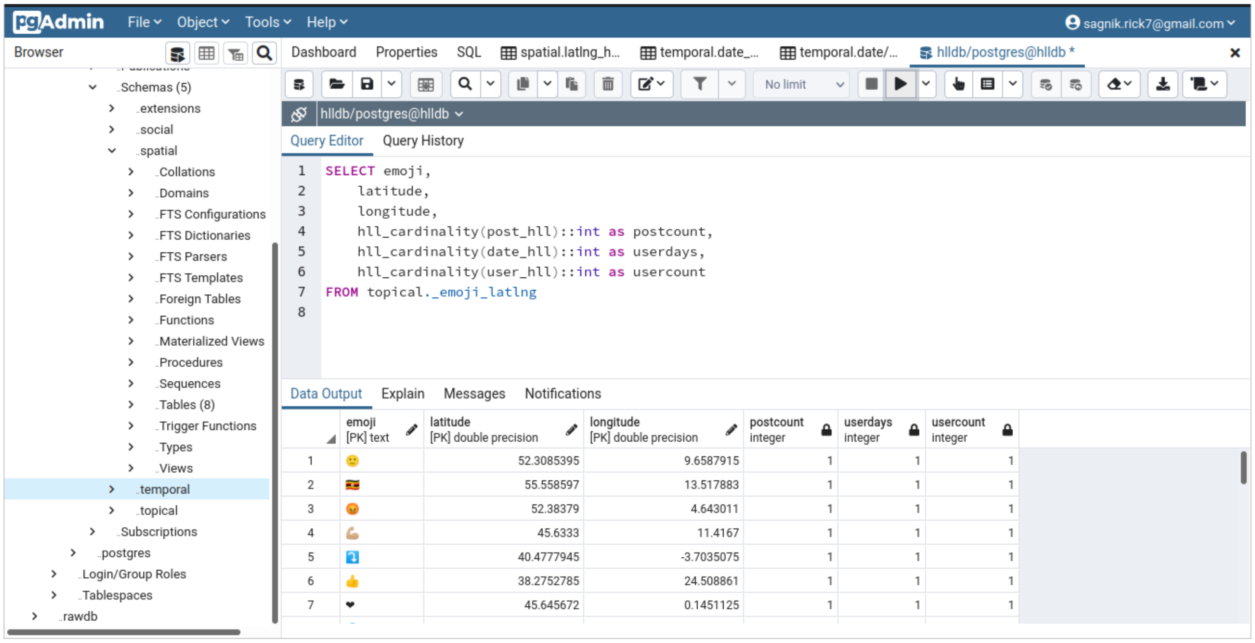
The HLL server stores the raw data in the HLL format. The conversion of the raw twitter data into the HLL format involves a schema tailor made for the raw dataset being used. The HLL data is divided into the facets of LBSN. Each facet holds the twitter data pertaining to that facet. For example, the query in Figure 6 shows us the contents of the temporal facet. Similarly, Figure 7 shows the result of a query in the topical facet.



```sql
SELECT date,
    hll_cardinality(post_hll)::int as postcount,
    hll_cardinality(user_hll)::int as usercount
FROM temporal.date
order by postcount desc;
```

	date [PK] date	postcount integer	usercount integer
1	2017-01-29	4417	2641
2	2017-01-30	3329	2088
3	2017-01-28	2441	1450
4	2016-06-20	2222	1353
5	2017-06-20	1961	1240
6	2018-06-20	1279	961
7	2017-01-31	1056	741

Figure 2 Query of the temporal schema



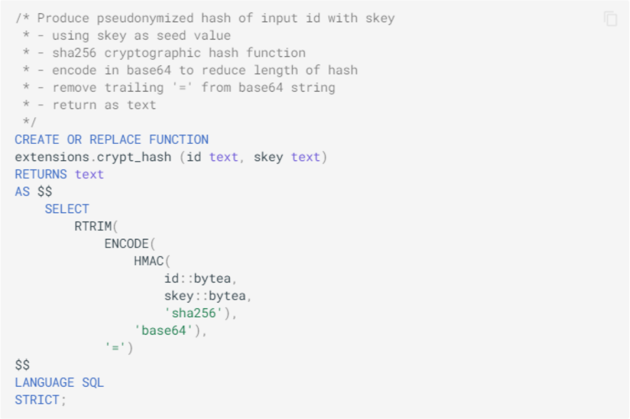
```sql
SELECT emoji,
    latitude,
    longitude,
    hll_cardinality(post_hll)::int as postcount,
    hll_cardinality(date_hll)::int as userdays,
    hll_cardinality(user_hll)::int as usercount
FROM topical._emoji_latlng
```

	emoji [PK] text	latitude [PK] double precision	longitude [PK] double precision	postcount integer	userdays integer	usercount integer
1	🙂	52.3085395	9.6587915	1	1	1
2	🇺🇬	55.558597	13.517883	1	1	1
3	😡	52.38379	4.643011	1	1	1
4	💪🏽	45.6333	11.4167	1	1	1
5	⤵️	40.4777945	-3.7035075	1	1	1
6	👍	38.2752785	24.508861	1	1	1
7	❤	45.645672	0.1451125	1	1	1

Figure 3 Querying the topical schema

The point to be noted in both the query examples is the fact that the topical and the temporal are very disconnected. Therefore, we do not have associated dates with the emojis. Using the metric user days we can estimate how many distinct users have used a particular emoji on a particular day. However, to estimate how many emojis have been used on one distinct day is not easily possible and is cumbersome.

One of the approaches to strengthen the security of the users along with the HLL implementation is geo-hashing. Like the HLL algorithm, geo-hashing was not originally meant for safeguarding user privacy, but this is a side effect of the algorithm. Geo-hashing reduces the accuracy of spatial coordinates by dividing the earth repeatedly into two halves. A decimal number indicates the degree to which the repeated sub-division will continue. POSTGIS has a geo-hash function which takes latitude, longitude, and desired accuracy (in the form of an integer) as inputs and returns geo-hashed coordinates. An example of such an implementation is shown in figure 8.



```
/* Produce pseudonymized hash of input id with skey
 * - using skey as seed value
 * - sha256 cryptographic hash function
 * - encode in base64 to reduce length of hash
 * - remove trailing '=' from base64 string
 * - return as text
 */
CREATE OR REPLACE FUNCTION
extensions.crypt_hash (id text, skey text)
RETURNS text
AS $$
    SELECT
        RTRIM(
            ENCODE(
                HMAC(
                    id::bytea,
                    skey::bytea,
                    'sha256'),
                'base64'),
            '=')
$$
LANGUAGE SQL
STRICT;
```

Figure 4 Geo-hash function in POSTGIS
Copied from *(A Basic Visual Analytics Example - LBSN Structure, n.d.)*

Since the exact coordinates are now not revealed, it adds a further layer of security to the twitter data.

As discussed in section 2, HLL data does allow for analysis through the union and intersection functions. The example involves finding co-occurring hashtags in tweets. The interesting point about the union function is that it counts only the distinct users in the sets. If on Day X there are 10 posts (post_count) and on Day Y is 6, then the union will not be a simple addition of 16. It will only be 16 if on both days X and Y there are unique users who have tweeted about the migration crisis. This is referred to as a lossless union.

The inclusion-exclusion principle forms the basis for the other analysis tool for HLL sets, intersection. Equation 4 and Equation 5 depict the principle. Here the || denotes cardinality the ∪ denotes union and ∩ denotes intersection. These principles can also be depicted graphically as in figure 6.

$$A \cup B| = |A| + |B| - |A \cap B|$$

Equation 4 Inclusion-Exclusion Principle for 2 sets

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$$

Equation 5 Inclusion-Exclusion Principle for 3 sets

In terms of the case study, the union and intersection function can be used to find the number of users who have used three different hashtags. First the cardinality of all the three individual sets must be calculated and then followed by the cardinality of the intersection set.

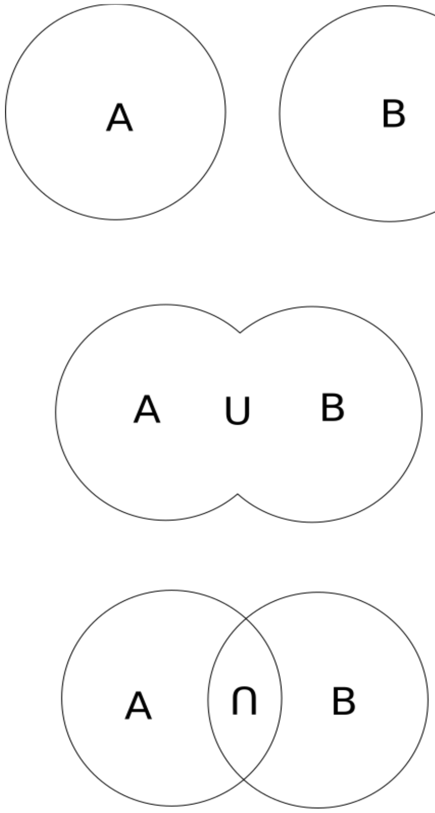
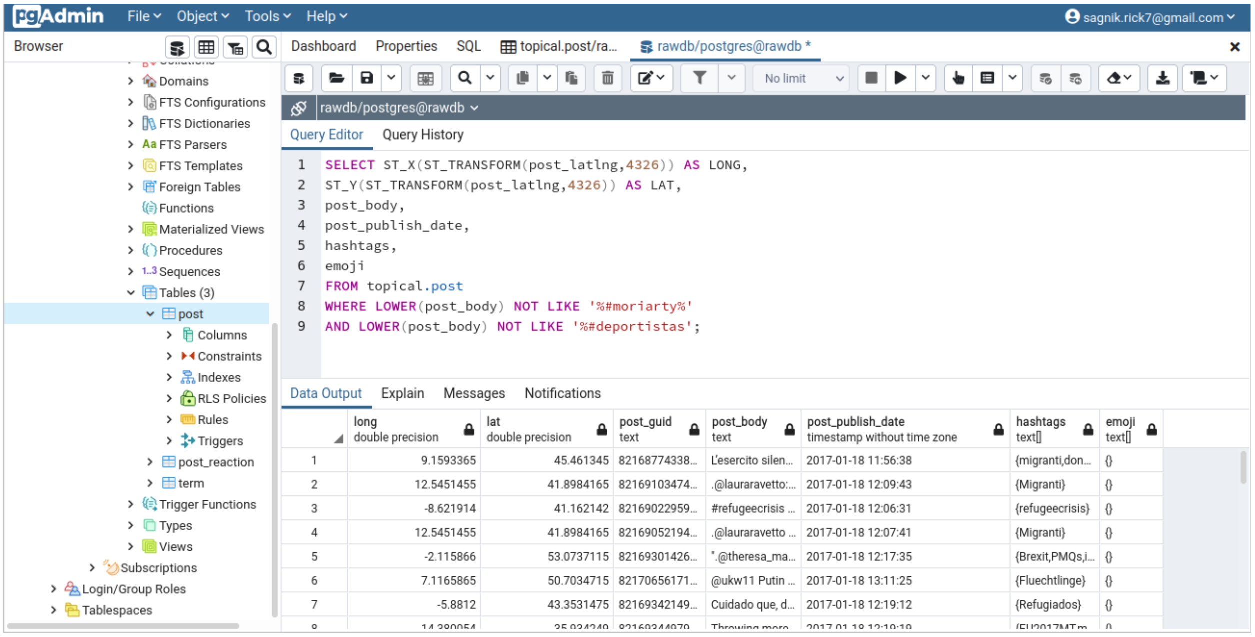


Figure 5 Graphical Representation of Union and Intersection

3.3.2 Raw

The raw data stored in the server came with several different columns. All the columns were not relevant to this work, and the relevance was decided based on the facets to be explored. In the query shown in figure 7, columns relevant to the spatial, temporal, topical and social facets are shown.

The query exhibits the use of the ST_TRANSFORM along with the ST_X and ST_Y functions which are POSTGIS functions for transforming POSTGIS geometry objects from the OGC Well-Known-Text (WKT) format into latitude and longitude points. The last line of the query also shows two more lines of filtering in English and Spanish. During the exploration of the raw dataset, it was found that the filtering performed in the initial stages was not adequate. The query 'moriarty' which relates to the BBC produced drama series Sherlock Holmes was most probably picked up because of having 'Moria' in the query. Moria was a refugee camp on the Greek island of Lesbos which closed in September 2020. Similarly, 'deportistas' (in Spanish) could have been picked up due to the presence of 'deport' (in English). However, it should be stressed that the number of irrelevant tweets, at worst, is not more than 2% of the total tweets.



```sql
SELECT ST_X(ST_TRANSFORM(post_latlng,4326)) AS LONG,
ST_Y(ST_TRANSFORM(post_latlng,4326)) AS LAT,
post_body,
post_publish_date,
hashtags,
emoji
FROM topical.post
WHERE LOWER(post_body) NOT LIKE '%#moriarty%'
AND LOWER(post_body) NOT LIKE '%#deportistas';
```

	long double precision	lat double precision	post_guid text	post_body text	post_publish_date timestamp without time zone	hashtags text[]	emoji text[]
1	9.1593365	45.461345	82168774338...	L'esercito silen...	2017-01-18 11:56:38	{migranti,don...	{}
2	12.5451455	41.8984165	82169103474...	.@lauraravetto:...	2017-01-18 12:09:43	{Migranti}	{}
3	-8.621914	41.162142	82169022959...	#refugeecrisis ...	2017-01-18 12:06:31	{refugeecrisis}	{}
4	12.5451455	41.8984165	82169052194...	.@lauraravetto ...	2017-01-18 12:07:41	{Migranti}	{}
5	-2.115866	53.0737115	82169301426...	".@theresa_ma...	2017-01-18 12:17:35	{Brexit,PMQs,i...	{}
6	7.1165865	50.7034715	82170656171...	@ukw11 Putin ...	2017-01-18 13:11:25	{Fluechtlinge}	{}
7	-5.8812	43.3531475	82169342149...	Cuidado que, d...	2017-01-18 12:19:12	{Refugiados}	{}

Figure 6 An example query for the dataset used while exploring raw data

The exploration of raw data followed mostly along the facets of LBSN to achieve the objectives outlined in section 2. The main workflow of analysis for all the facets in the raw data is shown in figure 11. A brief description of the steps involved is given below.

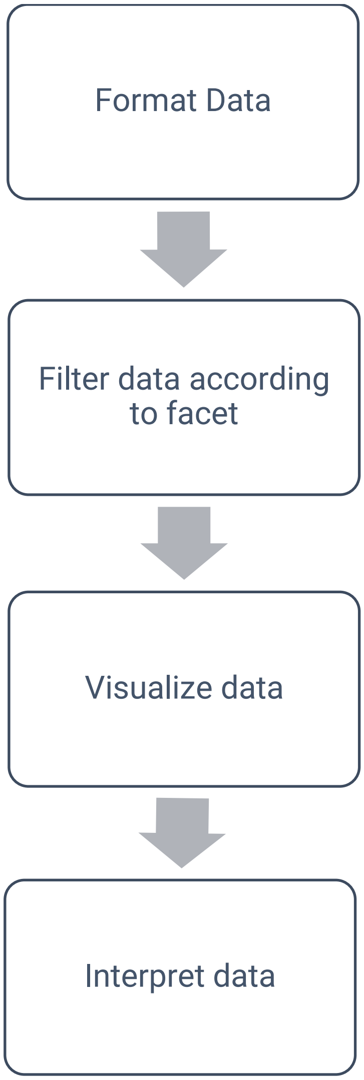


Figure 7 Workflow of raw data analysis

Formatting the data depends on the columns to be explored and on the Python library as well to be used. The bulk of the formatting was on the temporal facet as twitter timestamps are defined to seconds (YYYY-MM-DD HH:MM: SS). Using such timestamps for all the cases would not only clutter visualizations seeking an overview of the data but would also be a potential privacy vulnerability. Hence aggregating timestamps to coarser monthly or yearly units were required. Pandas python library had methods to deal with such temporal aggregation.

The other columns requiring extensive formatting would be the post_body, hashtags and emojis. These required the very standard formatting of string data, trimming blank spaces from the front and end of strings and changing all letters to lowercase. Tweets (post_body) were cleaned using the preprocessor library in preparation for sentiment analysis. This involved removing emojis, URLs, mentions, smileys and numbers.

Hashtags were cleaned by removing the # symbol and keeping the words of the hashtags. This tweet cleaning for sentiment analysis will be relevant for a later discussion on the results of sentiment analysis.

The next step of filtering the data is relatively straightforward. It involves filtering and slicing the data according to the sought-after results. For example, if one is interested in looking for the significantly active months of tweeting about the refugee crisis in Germany, there are two filters to be applied. The first step would be to single out the coordinates in Germany from the list of latitude and longitude data by means of a spatial join. Alternatively, one can also create a bounding box of latitude and longitude for slightly less accurate results. The next step would be to filter the aggregated monthly timestamps and count the rows for each month. Furthermore, typicalities of the frequent hashtags could be the following step in the analytical pipeline.

Data visualization would be dealt with extensively in detail in section 3.4. In brief, visualizing the data has been the result of the steps explained above. Since the facets themselves are interdependent on each other for effective evaluation, the visualizations must also relate to each other for proper interpretation.

The final and arguably the most challenging part of the entire workflow is the interpretation of the data. The challenge at this step mainly comes from the case study selected for applying the facets of LBSN, namely the migration crisis in the EU. As the dataset used for the thesis spans from 2016 till 2021, users were reacting to various events taking place during this time. Hence, fully understanding the events from the data also requires a thorough knowledge of the history of the crisis. This is further hindered by the multilingual nature of the dataset. This will be addressed extensively in section 4.3. For now, the workflow for data interpretation involves connecting the peaks on twitter data to events which took place during the past. In other words, an attempt was made to find causality, or perhaps it would be safer to say a correlation, for the tweets and events pertaining to the migration crisis. The data sources used to search for these events from the temporal, topical and spatial facets were news media like Deutsche Welle and The Guardian. Links found between twitter events and reports from the news were attempted to be placed within the broader scientific research.

3.4 Data Visualization

Visualizing the data was not only the result of exploring the results from each of the facets, but also the result aimed at as the final visualization of this thesis. The results of every facet required different visualizations. First, the author offers a brief overview on the visualizations involved before moving on to the combined visualization of all the facets. For reasons to be discussed in detail in section 4.1, visualizations of the HLL data would only be discussed under the spatial facet.

3.4.1 Spatial

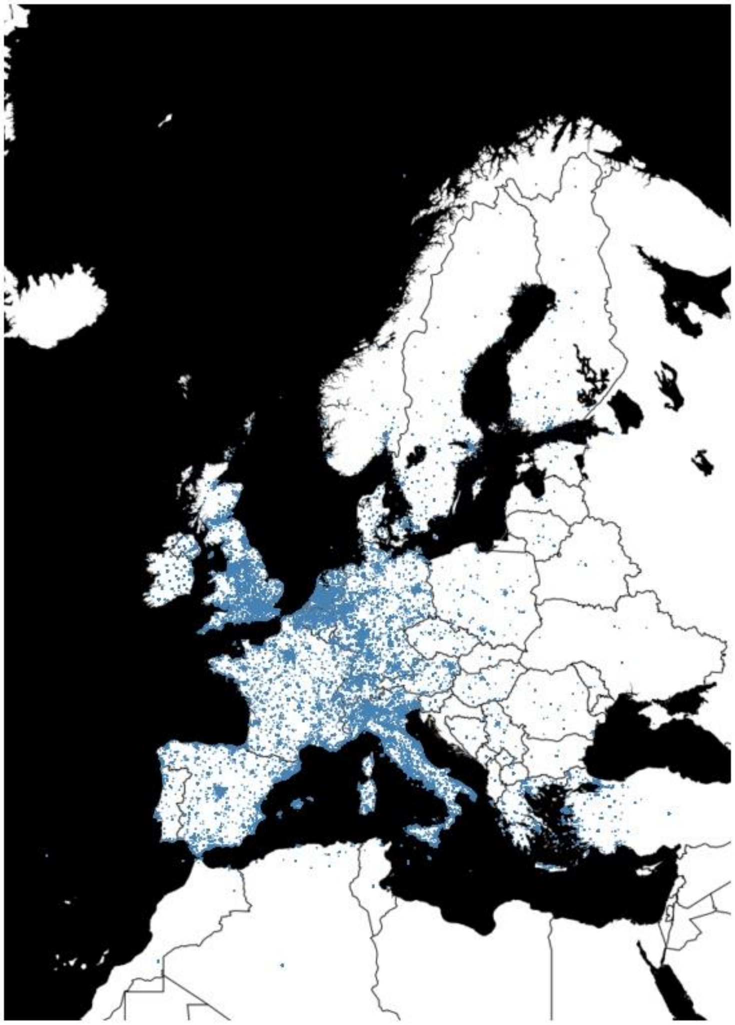


Figure 8 All tweets collected for the thesis

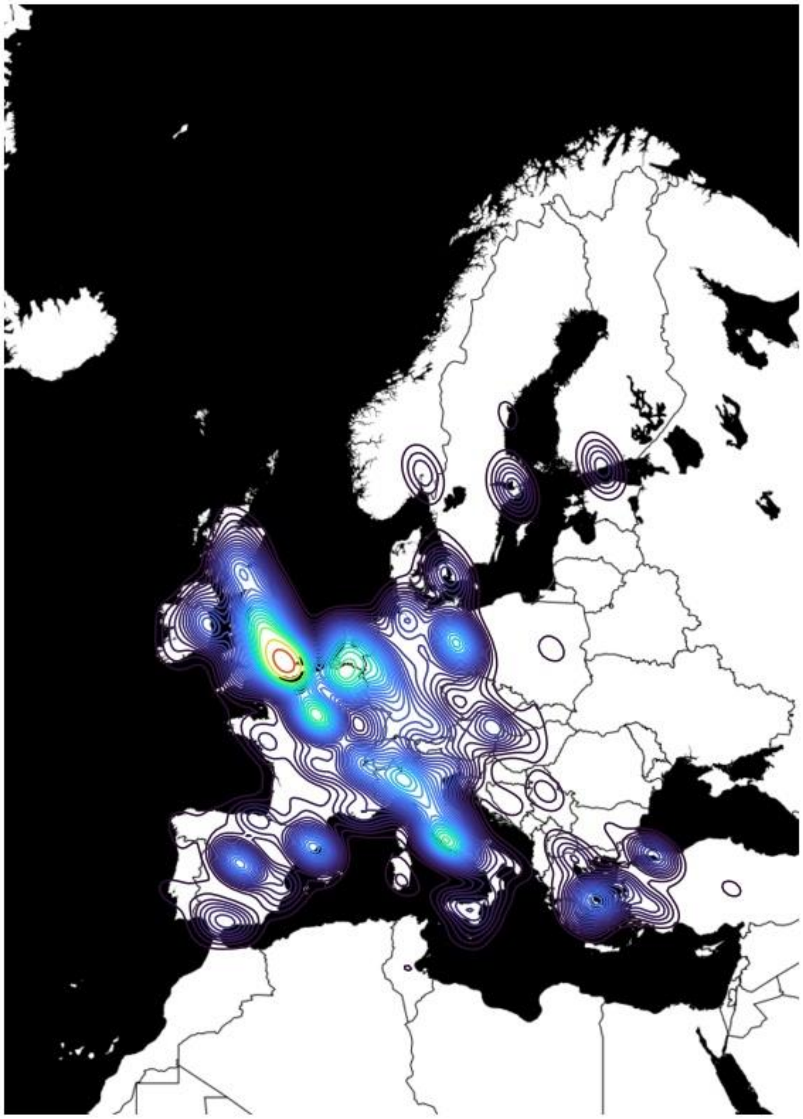


Figure 9 KDE of the tweets to identify clusters

The spatial facet was mostly visualized using maps. The kinds of maps used were naturally depending on the situation involved. Figure 9 depicts all the locations of the tweets used for this thesis. The basemap is derived from the spatial extent of the tweets using the python libraries geoplot (Aleksey Bilogur et al. 2019) and contextily .

Figure 10 shows a kernel density estimation (Rosenblatt 1956) to better visualize the clusters among the location of tweets. This was also achieved using the kdeplot function within the geoplot library. Figure KDE better shows the hotspots of tweets, which are mainly expected to be concentrated around densely populated urban areas (G. Almatar et al. 2020).

3.4.2 Temporal

There were two major kinds of visualizations used for the temporal facet. Figure 14 shows bar graphs showing the number of tweets across the timespan of the data, either yearly or monthly.

Figure 15 shows a line graph of trends which also shows the number of tweets; however, it also contains the number particular to every country. It is true that both these graphs are the result of similar data, the number of tweets at any given time. The usage of two kinds of visualizations was for the dashboard's usability. This will be further discussed in section 3.4.5.

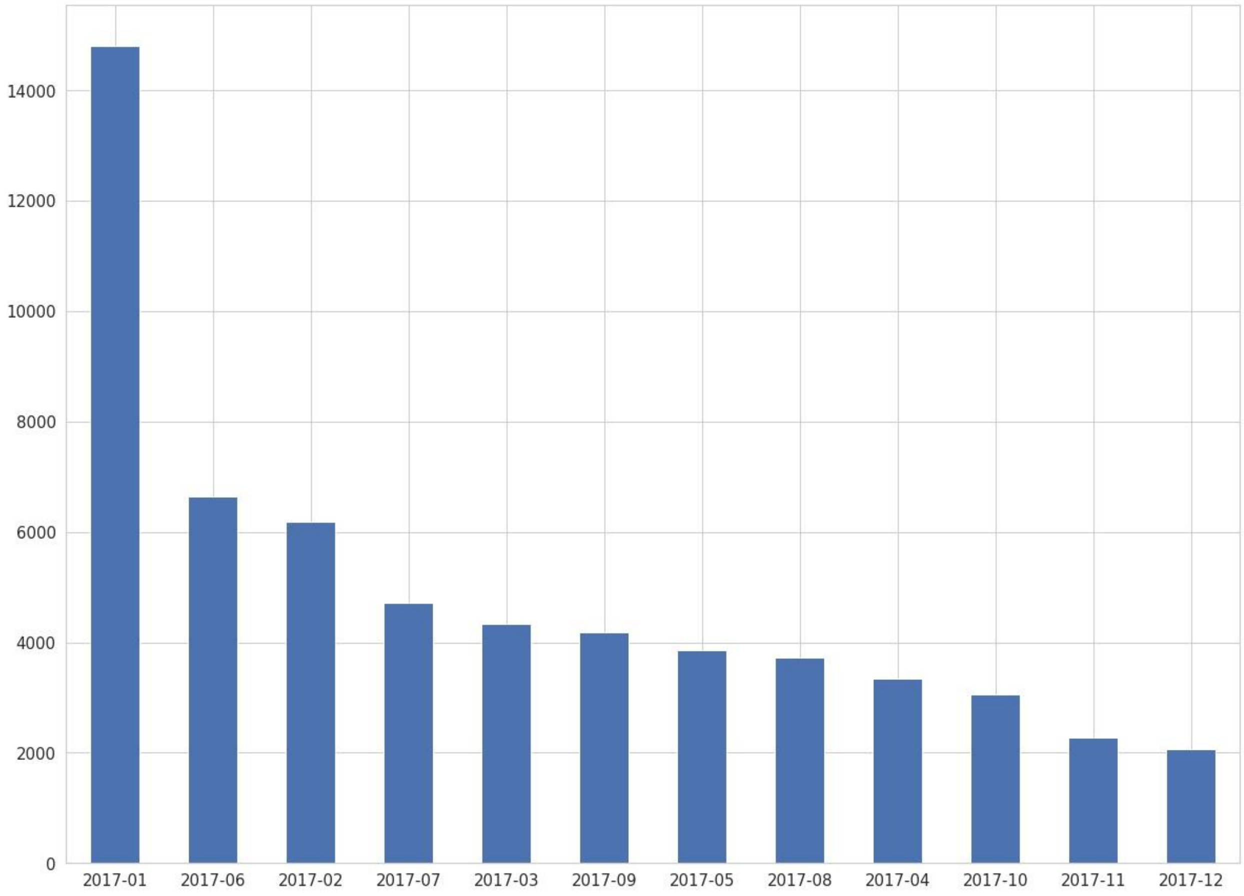


Figure 10 Number of tweets per month for the year 2017

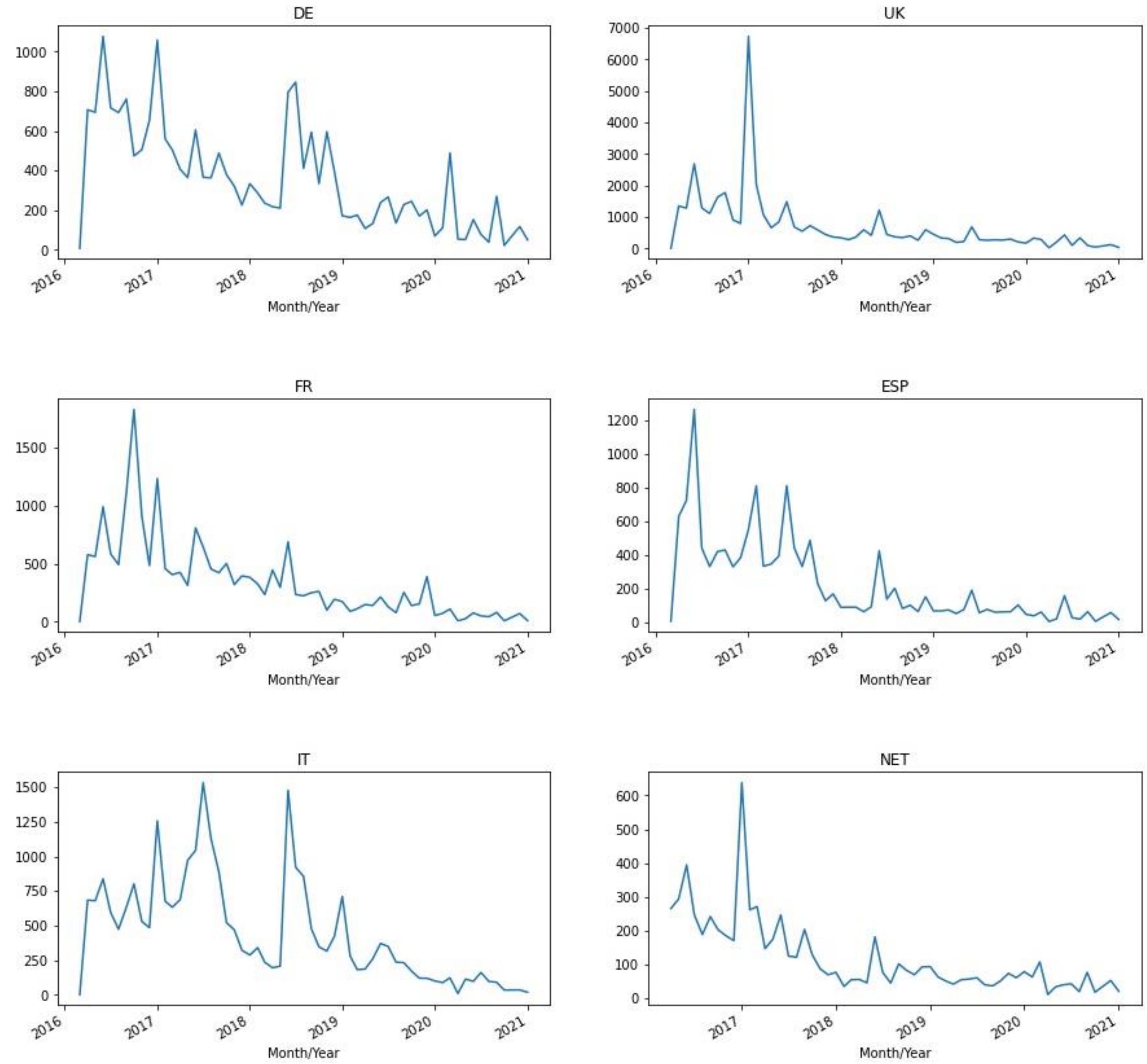


Figure 11 Tweet trends for various countries

3.4.3 Topical

Visualizations of the topical facet was primarily word clouds. The word clouds were used to quickly inspect the hashtags with the most frequencies within a certain spatio-temporal bounds (hashtags used during the month of January 2017 in Germany). There was also the use of word clouds to find co-occurring hashtags for a particular hashtag. As an example, to make a wordcloud of the co-occurring hashtags of the hashtag *migranti* would be to find all the posts with the hashtag *migranti*, extract all the hashtags, remove the hashtag *migranti* and then create the word cloud using the wordcloud python library. The result is shown in figure 13.

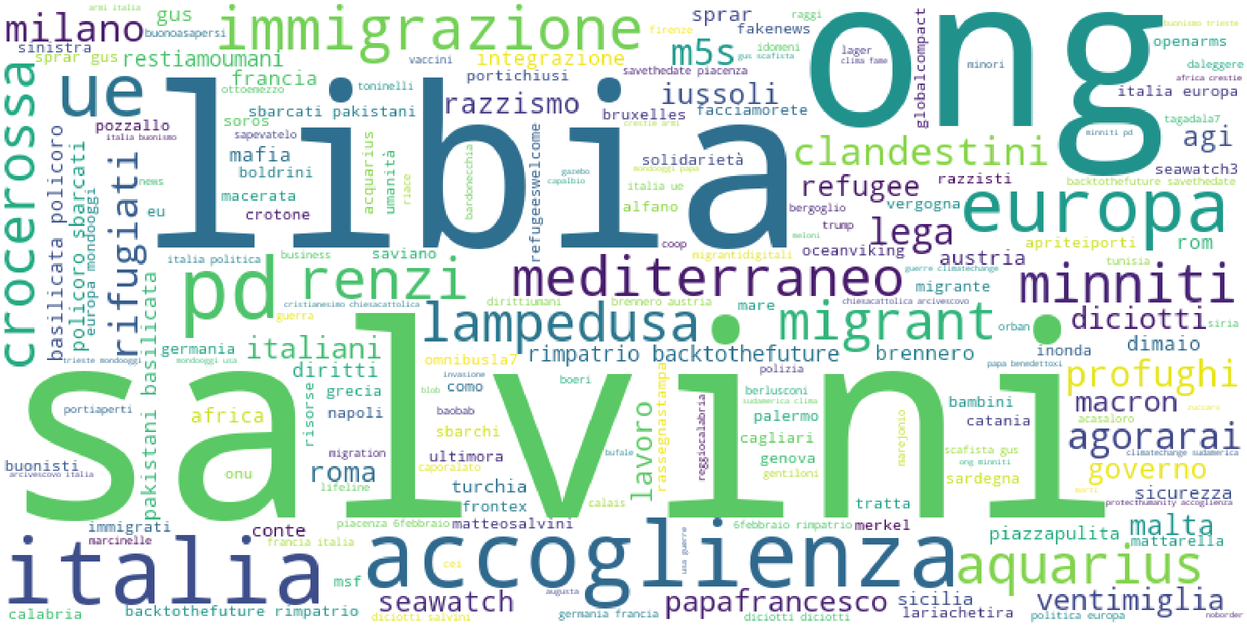


Figure 12 Wordcloud of Co-Occurring hashtags with *migranti*

3.4.4 Miscellaneous

Before moving to explicate on the dashboard, there are two other cases which should be addressed. The first one being visualizations of HLL datasets and the second one being typicality.

3.4.4.1 HLL data visualization

There was only one major visualization of the HLL data for this work. This visualization is of usercount in Germany using HLL data, shown in figure 14. The objective of this visualization was to further enhance the security of users while showing the spatial distribution of the metrics. This is an HLL data alternative of the raw data KDE map. This map requires making 250 km x 250 km grids within the latitudinal and longitudinal extent of Germany. Each grid created, is thus a polygon. These polygons are assigned points (derived from geohashed latitude and longitude data associated with HLL dataset) using a spatial join. Then the resulting geopandas dataframe is cleaned of NaN (not a number) values and plotted. The classification scheme used for this was head tail breaks as this is the most suitable classification for highly skewed datasets (Jiang 2013) .The map tile used for this map was the Stamen Toner Hybrid which gives the labels for a number of major and minor cities.

A lingering issue remained with this visualization. After the spatial join which assigned values of usercount to the grids created, the usercount values could not be added to a single value. As a result, every grid had multiple usercount values, instead of just one value. This affects the visualization because the legend displays the highest usercount value to be 284. However, none of the grids, corresponds to the color indicated for the value of 284. The python library probably uses a median or mean value from the different usercount values and colors the grids accordingly. The result still shows the regions where most users have tweeted from, but the differences between the grids should be more than what the map currently shows.

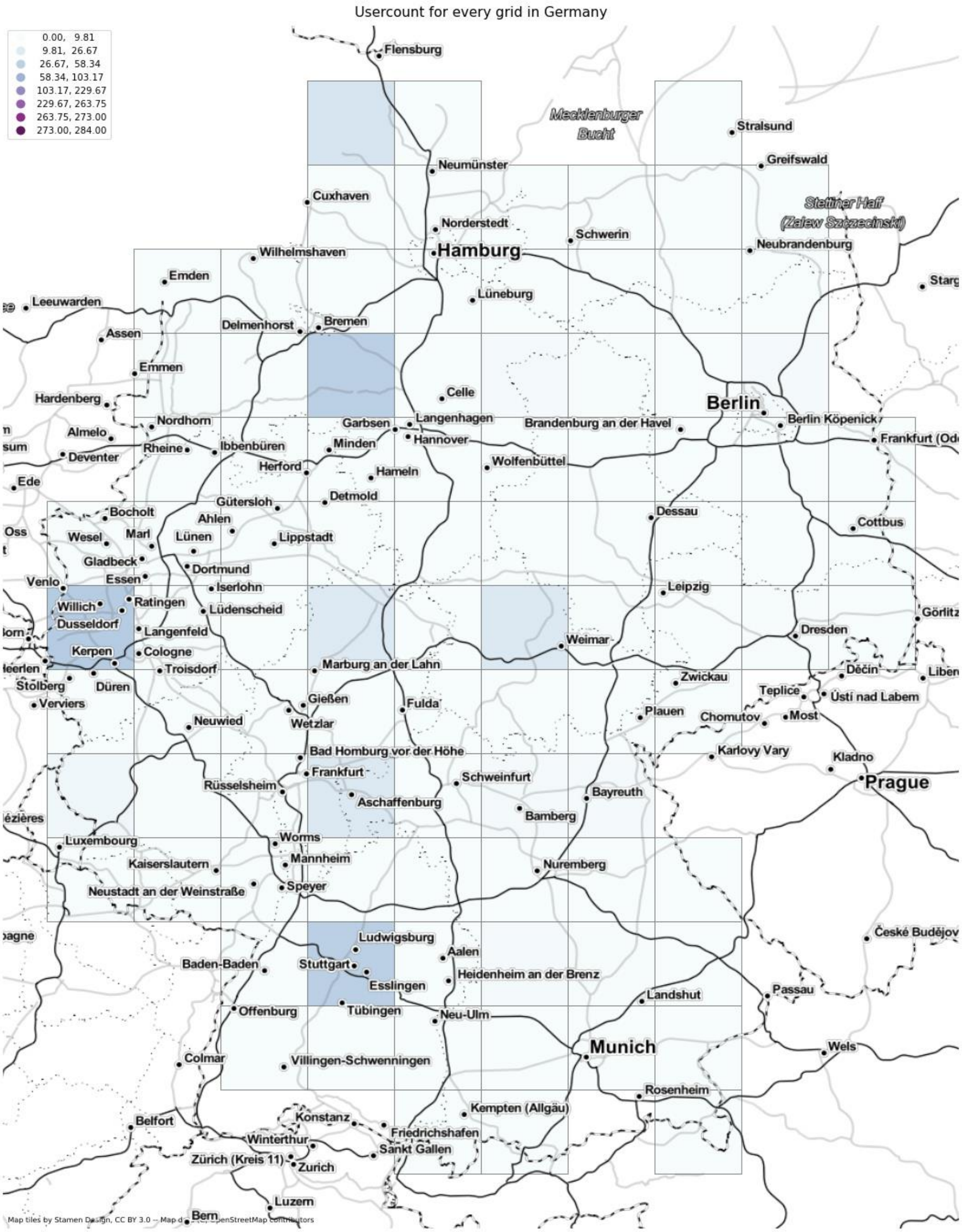


Figure 13 usercount using HLL data for Germany

3.4.4.2 Typicality

For this work, typicality measured has only been of hashtags based on various temporal subsets. Ideally, this would be categorized as both temporal and topical facet. The visualizations used were grouped bar charts shown in figure 15. However, the dashboard features stacked bar charts and normal bar charts.

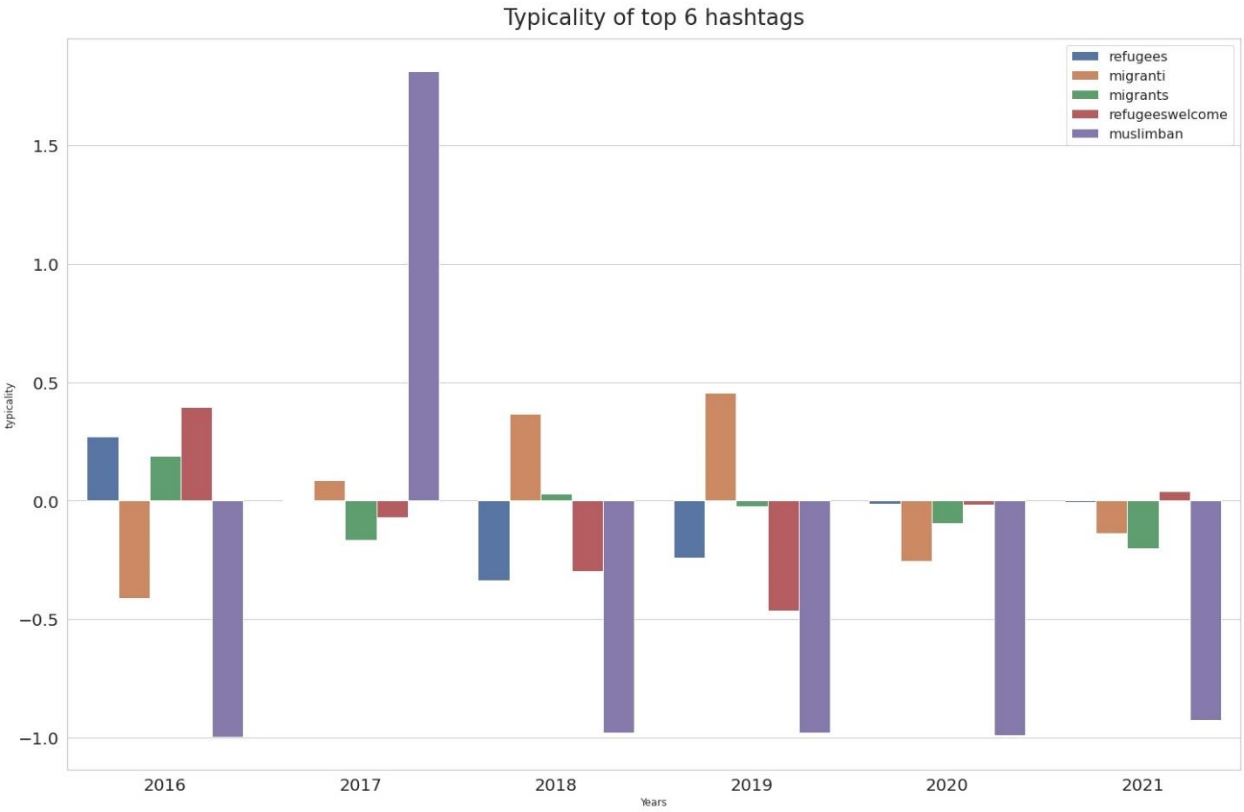


Figure 14 Grouped bar chart illustrating the typicality of some popular hashtags

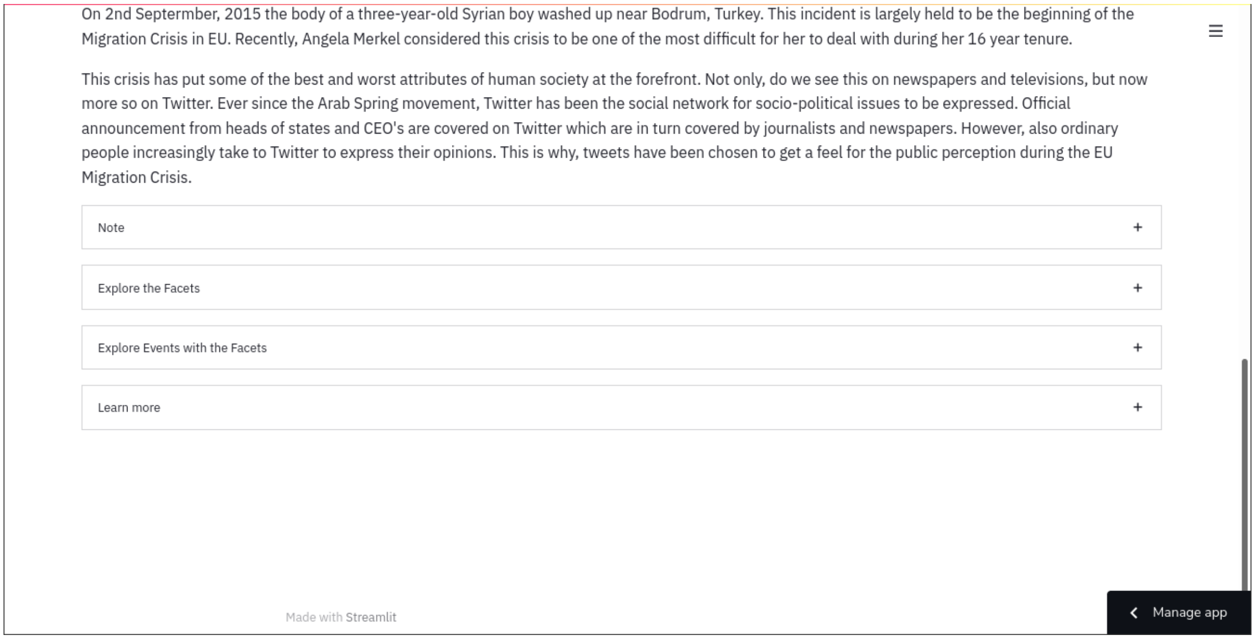
3.4.5 Dashboard

Every facet on its own only shows a part of the data. It is when the facets connect, through a comprehensive visualization, a holistic picture of the dataset is revealed. This was the primary reason for choosing the dashboard as the final visualization. A secondary reason was also the interactivity that the dashboard offers.

The dashboard was made using a python library called Streamlit. The library allows for making easy interactive dashboards and more importantly allows seamless hosting capabilities using Github. Another advantage of Streamlit is the supported mapping library, pydeck.gl. Pydeck.gl is a python library based on the Javascript deck.gl library which makes creating high level spatial visualizations simple.

There were two separate dataframes created out of the original raw data. One was used for the Altair (VanderPlas et al. 2018) typicality chart under the Topical subheader. The other was modified and used for all other visualizations in the entire dashboard. There were also three images used in the dashboard: the image of refugees fleeing Moria and the two timelines. The goal was to make the dashboard into a narrative visualization while also allowing for data exploration and the fullest interactivity.

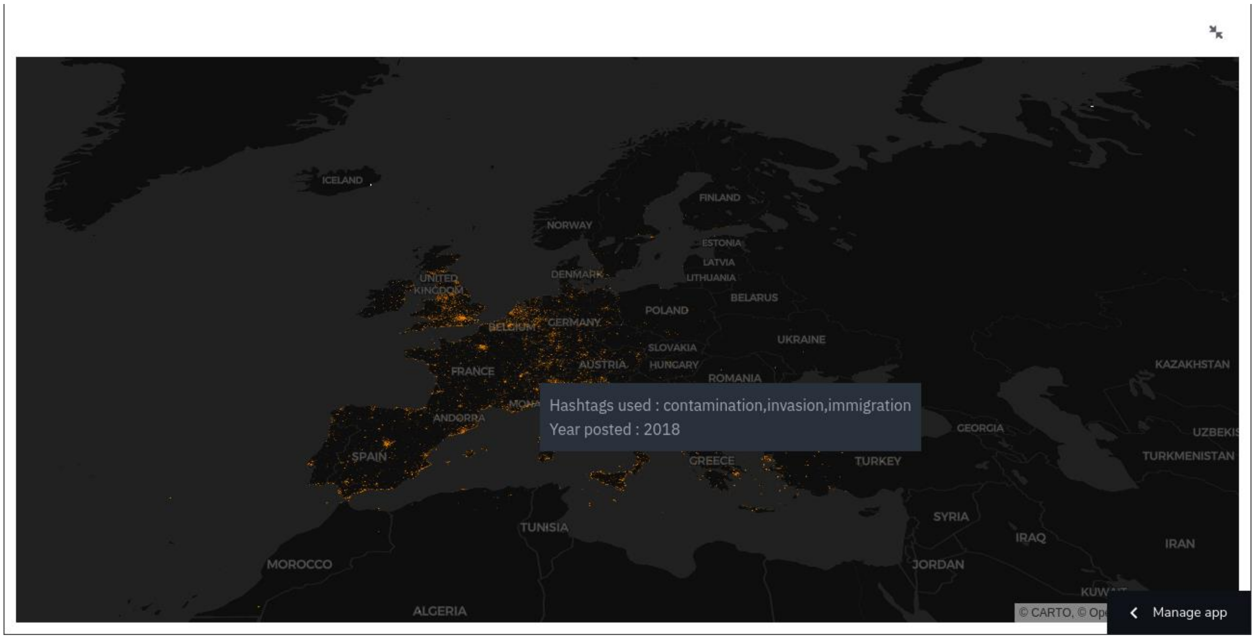
Incorporating elements of narrative visualization and allowing for interactivity were achieved using various methods within the Streamlit library. Figure 16 shows the principles of abstraction and elaboration implemented using the expanders. Figure 17 shows the inclusion of information from the topical and temporal facet within the spatial facet using tooltips. Figure 18 illustrates the use of selection boxes to filter the data and display results according to the user's choice.



On 2nd Septermber, 2015 the body of a three-year-old Syrian boy washed up near Bodrum, Turkey. This incident is largely held to be the beginning of the Migration Crisis in EU. Recently, Angela Merkel considered this crisis to be one of the most difficult for her to deal with during her 16 year tenure.

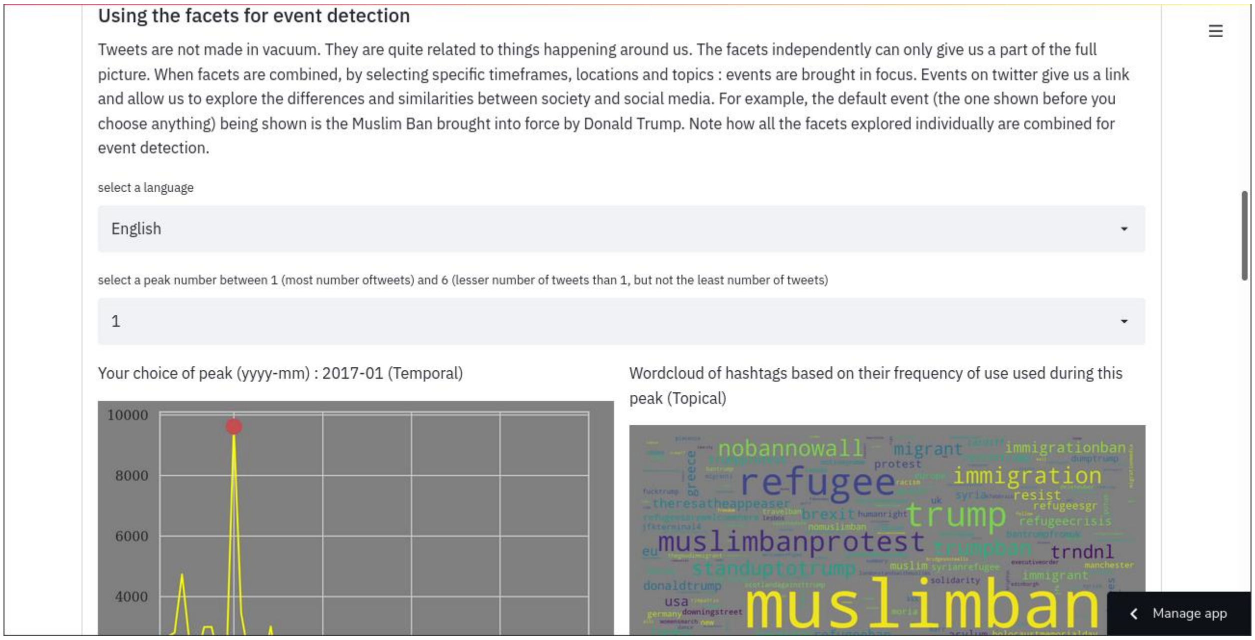
This crisis has put some of the best and worst attributes of human society at the forefront. Not only, do we see this on newspapers and televisions, but now more so on Twitter. Ever since the Arab Spring movement, Twitter has been the social network for socio-political issues to be expressed. Official announcement from heads of states and CEO's are covered on Twitter which are in turn covered by journalists and newspapers. However, also ordinary people increasingly take to Twitter to express their opinions. This is why, tweets have been chosen to get a feel for the public perception during the EU Migration Crisis.

Figure 15 Principle of abstraction and elaboration enacted through Streamlit expanders



Hashtags used : contamination,invasion,immigration
Year posted : 2018

Figure 16 Tooltip with contextual information for the spatial and temporal facet



Using the facets for event detection

Tweets are not made in vacuum. They are quite related to things happening around us. The facets independently can only give us a part of the full picture. When facets are combined, by selecting specific timeframes, locations and topics : events are brought in focus. Events on twitter give us a link and allow us to explore the differences and similarities between society and social media. For example, the default event (the one shown before you choose anything) being shown is the Muslim Ban brought into force by Donald Trump. Note how all the facets explored individually are combined for event detection.

select a language

English

select a peak number between 1 (most number oftweets) and 6 (lesser number of tweets than 1, but not the least number of tweets)

1

Your choice of peak (yyyy-mm) : 2017-01 (Temporal)

Wordcloud of hashtags based on their frequency of use used during this peak (Topical)

Figure 17 Allowing filtering for information exploration

4. Results and Discussions

4.1 Objective 1 (HLL)

HLL datasets for twitter data have both significant benefits and a few challenges. It must be noted that the conclusions from this work should not be taken as definitive statements. Working with HLL data is still at its nascent stages and it is recommended to use the raw data format along with the HLL format for analysis. Therefore, both the benefits and challenges described, should be indicative for further research and development using multiple and varied case studies.

In the context of data analysis and this work, the HLL format allows for accurate quantitative analysis. For example, the number of posts with the hashtag *refugees* from the raw dataset was 16,451. For the HLL dataset, the number stands at 16,441. Hence the error for this case is negligible and makes barely any difference in large scale datasets like the twitter migration data. The union and intersection function also gives leeway for qualitative analysis as demonstrated in the appendix (Carto_Master_Appendix/Scripts/HLL_auto.ipynb). It is also possible to include other facets, like latitude and longitude information which is also shown in the same Jupyter Notebook.

For the union and intersection functions, there are also cardinality estimation functions or libraries readily implemented in PostgresSQL or Python respectively. The python library is slower compared to the Citus HLL on PostgresSQL, but this can be mitigated by querying the PostgresSQL database directly through the python script using psycopg2. From there one could use the entirety of python libraries for visualization if one chooses not to use the materialized views on PostgresSQL. Hence once setup, the workflow is smooth and effortless.

However, there remains a few reasons for which the author chose to rely on the raw dataset rather than the HLL data for analysis and interpretation. For the time being, transforming the raw dataset into HLL requires quite a bit of effort and is cumbersome. The transformation of the data requires a mapping which converts columns from the raw dataset into columns which are HLL-ready. As a result, a few of the columns from the raw dataset are lost. For example, the raw data contains a column called input_source which could be used to filter bots from real users. This column was not present in the HLL data, potentially losing the ability for this filtering, if required.

This transformation also separated emojis and hashtags into two separate tables. In fact, the HLL format seemed most disruptive within the topical facet. As emojis and hashtags are separated not only from each other but also from the original post, a lot of the contextual information is lost. This makes a further qualitative analysis of the dataset impossible. This challenge persisted also between other facets.

The secondary reason for prioritizing the raw dataset over the HLL is also to minimize the loss of contextual information. This was mainly applicable to the explorations in the topical facet and how the facets can be used to characterize an event. According to Dunkel et. al. (2018) an event is described as an object of the temporal facet. Following this definition would have been consistent with the HLL format but would have lost the spatial and topical features. For the case of the migration crisis, this would have been very misleading.

Consider the month of June 2018. If one looks at the ongoing events in Germany (Peak 3 on the dashboard), they will find the Merkel government in crisis over Horst Seehofer's resignation. However, at the same time in Italy (Peak 2 on the dashboard) the then deputy prime minister Matteo Salvini announced the closing of ports to ships carrying "illegal immigrants." This example illustrates the difference that both the spatial and the topical facet makes in interpreting the data.

It should be reiterated that the HLL format does perform its fundamental function of being privacy aware extremely well. The cyyptographically hashed hexadecimal representations are formidable safeguards against privacy breaching attacks. It becomes a daunting task to figure out which post_guid from the raw dataset would represent the hashed string from the post_hll column in the HLL dataset. By further increasing the spatial granularity using grids for representing spatial distribution of the metrics, makes it further difficult to gauge exact locations of the posts or the users.

For these benefits and challenges the author feels that the HLL format will perform better with real time data rather than data from the past. This is because using real time data, the researchers would be aware of what kind of data they are working with and the event related to the data being searched. In the case of this work, the author had to find the events from the tweets. The latter case might not be suited for the HLL format because of the dissociated facets within the dataset.

4.2 Objective 2 (Visualization)

Before discussing the dashboard in detail, the author wishes to clarify the reasons for a choice made while building the dashboard. For visualizing the spatial facet, the author chose to used language for filtering instead of the geo-tagged location. This was because of choosing the boundaries of countries, instead of the language, reduced the number of tweets by approximately 16,000. Moreover, the trends for tweets during the peaks detected, remained almost the same. This can be seen in figures 19 - 21

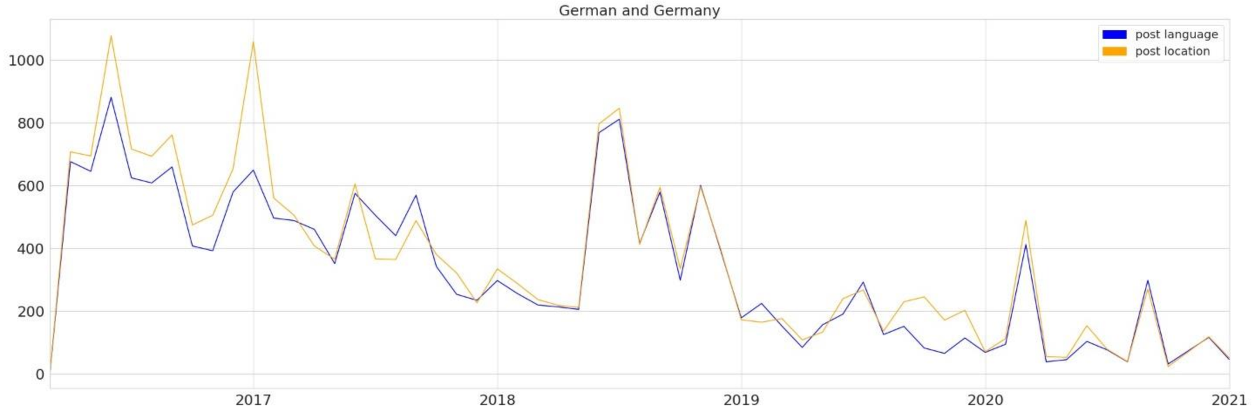


Figure 18 trendline for tweets in German and Germany

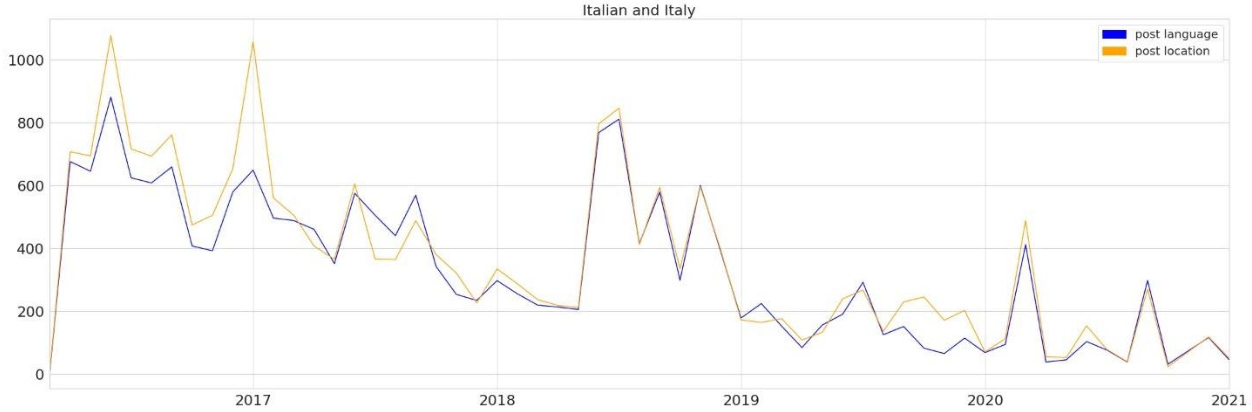


Figure 19 trendline for tweets in Italian and Italy

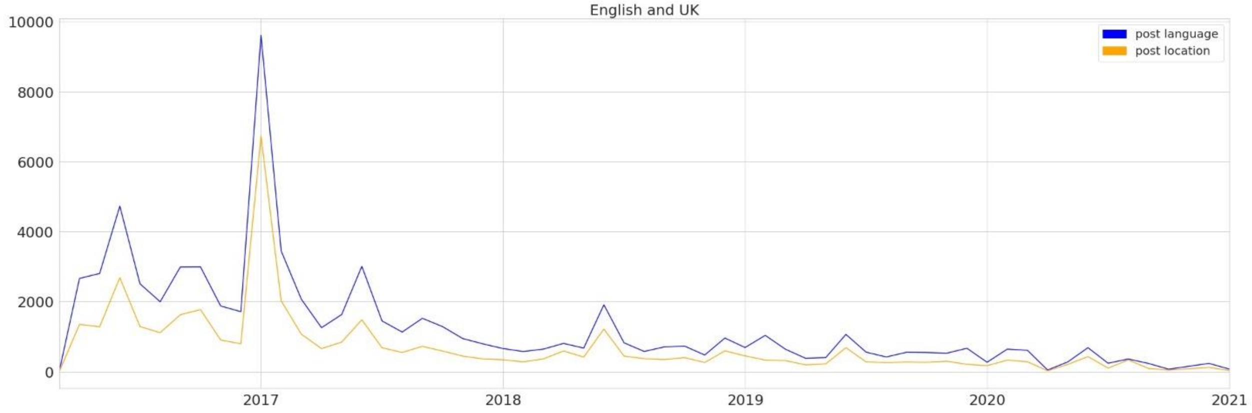


Figure 20 trendline for tweets in English and United Kingdom

In this dashboard certain elements of narrative visualization were incorporated, namely situatedness and event-sequencing are used (Caracciolo 2009). This was primarily included to remind users that what the analyses of tweets displays is not an abstract online activity being referred to. In short, this was the author's intentional choice to remind users of the predicament of millions of people. Along with this, narrative visualizations have been shown to improve cognition in information rich scenarios (Dove and Jones 2012) .The context and timeline were created with the photograph taken by Petros Giannakouris and the opening paragraph placing the users in the timeline of the crisis.

However, it must be stated that other elements of narrative visualization were not included. One of the reasons was the fact that twitter data does not present the entire reality of the refugee crisis. For a complete immersive narrative visualization, it is important to include other data sets specifically because the data collected was in the major European languages, where there is no representation of the plight of millions of refugees. Further data from traditional sources like government bodies for influx of refugees, opinion polls would be required for a compelling immersive experience for the user. There is also the issue of a disconnection in the dashboard. The charts and graphs pertain to the reactions expressed by the residents of Europe. But the opening picture refers to the crisis faced by the refugees. These two are in direct contrast with each other and could mislead the user's narrative experience. But, without a proper evaluation with a user group, this would only be a hypothesis and not a proven fact.

The dashboard made does not fit into the traditional description of a dashboard. A dashboard is defined as a system used to visualize data for decision making (Few 2006). This dashboard is close to a “pull” dashboard, as described by Janes A. et al. (2013). But it's utility as a system for decision making would be in question. Also, for the reasons already described the dashboard does not fully incorporate every element of a narrative visualization. One of the main reasons was that a user group was not defined for this dashboard. It can be said that intended audience would be non-experts who are only interested in understanding the data. However, this was not strictly followed during the designing process. Without a specific user group and evaluation methodology, it would be difficult to properly judge the effectiveness of the visualization in accordance with the principles of User Centered Design (Rubin and Chisnell 2008) .

To summarize this section, it can be stated that the dashboard falls short of being a dashboard in the strictest definition of its functionality. It also falls short of being a complete narrative visualization. The viability and effectiveness of such a design will be considered for future works.

4.3 Objective 3 (Data Interpretation)

In this section, a description of findings and observations within the dataset will be given. The section is divided into 4 different subsections and is concluded with an overall impression about the findings.

4.3.1 Sentiment Analysis

The reasoning and need for sentiment analysis were to look at the hashtags and observe the sentiment scores that was given to them. This was simple for English tweets. Figure 22 shows the scores for the hashtag *refugees* throughout the temporal facet of the dataset. The area in blue shows the standard deviation of scores. A score below 0 indicates negative sentiments, 0 indicates neutral sentiments and above 0 indicates positive sentiments.

The sentiment analysis was performed using VADER-Sentiment Analysis python library which is specifically attuned for sentiment analysis of social media texts (C. Hutto and Eric Gilbert 2014). This effectively means that the tweets did not have to be pre cleaned as described in section 3.2.2. However, this library was not available for other languages in the dataset. It was a problem using multiple libraries for each language as there would be no common basis for comparison of the results.

On top of that, there is the problem of cleaning the tweets. Most of the sentiment analysis libraries are catered towards analysis of normal texts, not social media posts. It is a daunting task of converting tweets into texts for sentiment analysis. It is rarely the case that removing the # symbol from hashtags (after removing number, URLs, etc) would turn the tweet into a proper sentence in English. Of course, this would be even more difficult for other European languages.

This would require training a classifier with cleaned tweet data and then implementing this classifier to the twitter dataset which was beyond the scope of this work and hence was not pursued further.

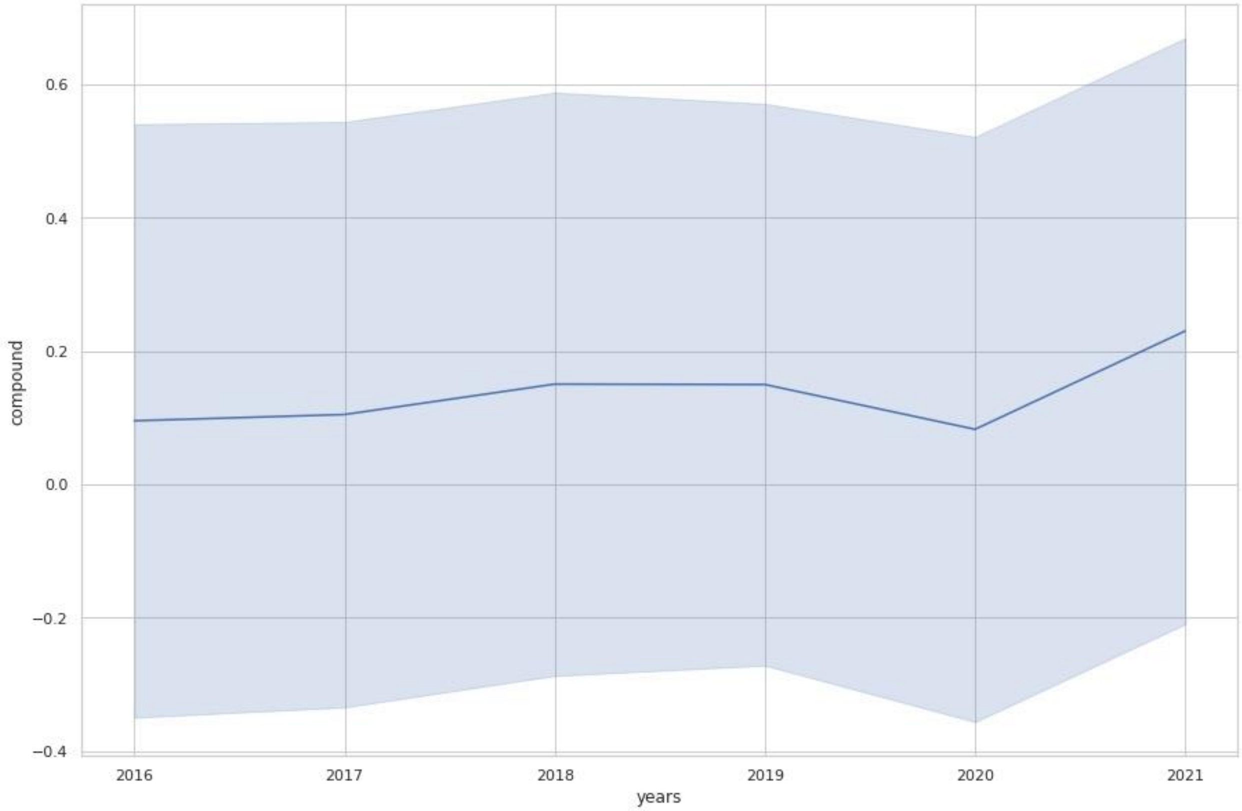


Figure 21 The sentiment scores for the hashtag *refugee* from VADER

4.3.2 Emojis

This work was initially intended to consider emojis for analyzing public reaction but of the 170,000 tweets only approximately 12,000 tweets had emojis with them. In light of this, extensive work was not performed with emojis.

The most popular emoji used was the 💓 emoji and the second most popular was the 😂 emoji followed by 🙏 emoji. According to the website emojipedia.org, these are some of the most popular emojis used on twitter in 2019. Therefore, it seems that even in the context of the migration crisis, the pattern of emoji usage is consistent with the larger emoji usage on twitter.

An unexpected find was the hashtag having the highest number of emojis. This hashtag was *refugeeswelcome*, even though by the number of posts with this hashtag ranked fourth (Table 3). The most popular emoji used with *refugeeswelcome* was the 💓 emoji. Other emojis used with *refugeeswelcome* are 💖, 🙌, ❤. Table 1 shows some other hashtags with the popular emojis.

Hashtags	Emojis
refugeeswelcome	💓, 💖, 🙌, ❤
migrants	→,●, Flag:France, 😧
refugees	💓, Flag:Greece, 🌍, 💗
muslimban	😂 , ▯, Flag:USA, 😧

Table 3 Hashtags and the 4 most used emojis along with them

From the above table *refugeeswelcome* seem to have overwhelmingly supportive emojis, at least when it's compared to *migrants* where the emojis cannot be classified into a particular affective category. However, further in-depth analysis of the context of tweets is required to draw any conclusions, which can be easily demonstrated by the following case.

The American flag emoji has been used by far-right groups in America to tweet against the Black Lives Matter movement (Alfano et al. 2021) . Here, *muslimban* has the American flag emoji associated with it, making the association with American far-right groups quite unlikely due to the context of this case study. On the other hand 🌍 emoji used with the *refugees* hashtag does indicate a spatial correlation as other globe emojis (🌎,) were not used.

Both the above cases do indicate that emojis are very ambiguous and context dependent (Hauthal et al. 2021). The low count of emojis in this dataset also did not build confidence in these observations. It is therefore prudent to take these observations as indicators to further explore the dataset.

4.3.3 Hashtags and Events

The entire dataset contains a total of 68,216 hashtags. A brief statistical summary of the hashtags is given in Table 2

Statistical Measure	**Result**
Total count	68216.00
Mean	6.538188
Standard deviation	132.996914
Minimum	1.00
First quartile	1.00
Median	1.00
Upper quartile	2.00
Maximum	16,456.00

Table 4 Summary of statistical measures of the hashtags

The table clearly shows that using a hashtag more than once is an exception within this dataset. Out of the 68,216 hashtags 23,632 hashtags have been used more than once and 14,481 hashtags have been used more than twice in the five years of the span of the dataset. The ten most popular hashtags are shown in table 3.

Hashtags	**Counts**
refugees	16,456
migranti	14,233
migrants	12,282
refugeeswelcome	11789
muslimban	10789
migration	8852
immigration	7454
refugee	4562
refugiados	3756
immigrazione	3082

Table 5 Most popular hashtags

On further exploring these hashtags along the temporal facet, it is possible to detect events. For this purpose, typicality can be used, by taking a temporal sub-dataset to see which hashtags were typical throughout the years. This is shown in figure 23

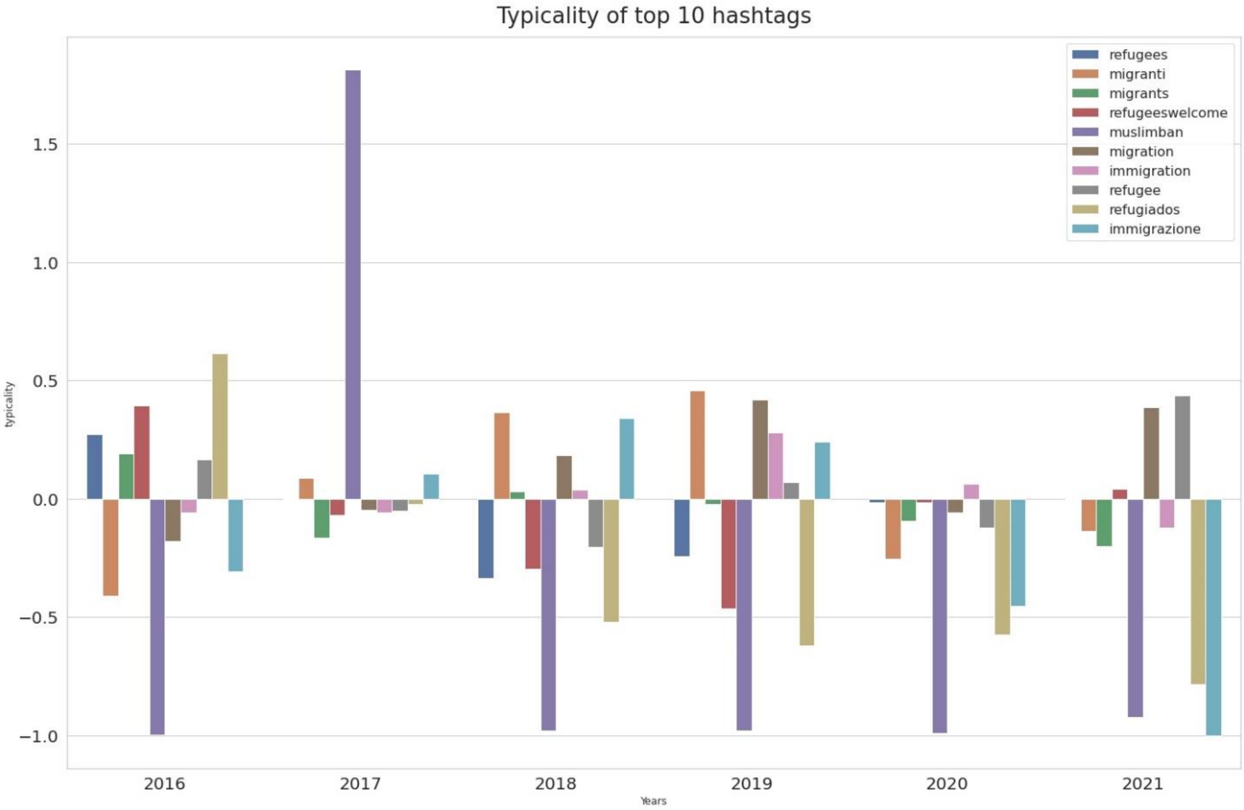


Figure 22 Typicalities of the 10 most popular hashtags

From figure 23, it can be concluded that hashtags like *migranti* and *immigrazione*, which are in Italian, correspond to events occurring in Italy. It is also the same for the Spanish hashtag *refugiados.* This conclusion is further supported by the trend lines of tweets in Italian and Spanish. The peaks seen in the trendline match quite well with the typicalities of the hashtags. The Spanish trendline is shown in figure 25 and the Italian in figure 24.

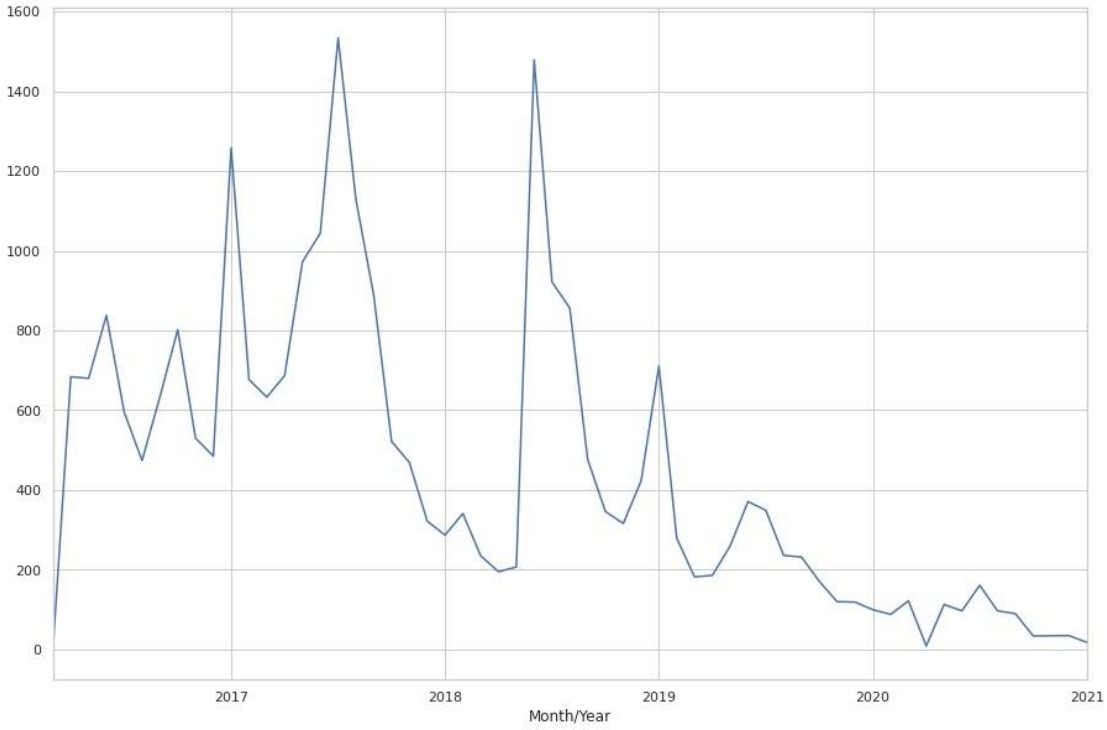


Figure 23 Trendline of tweets in Italian

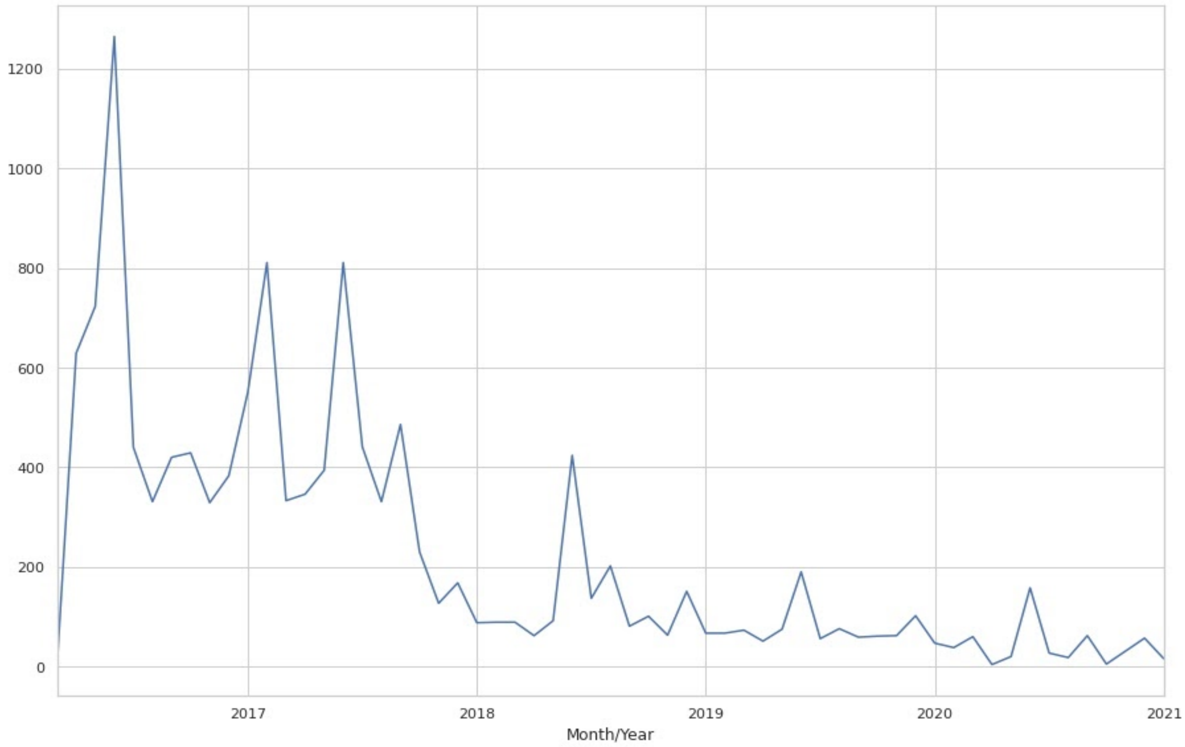


Figure 24 Trendline of tweets in Spanish

However, using the hashtag typicality to detect specific events has its limitations. For hashtags like *migration and immigration*, which are spelt in the same way for German and English. Many posts in languages other than English, also use English hashtags like *refugees* and *refugeeswelcome*. The problem also occurs when most hashtags are very general. The hashtag *refugee* can have tweets ranging over a wide range of topics, related to the migration crisis or even otherwise. Hence to overcome these problems, the co-occurring hashtags of the primary hashtag could be used once the temporal window for an event has been established from the trendline.

From figure 26 it is visible that in 2017 the hashtag *muslimban* was highly typical. It is also clear that this hashtag is atypical for the other years. Figure 27 shows the co-occurring hashtags along with *muslimban*.

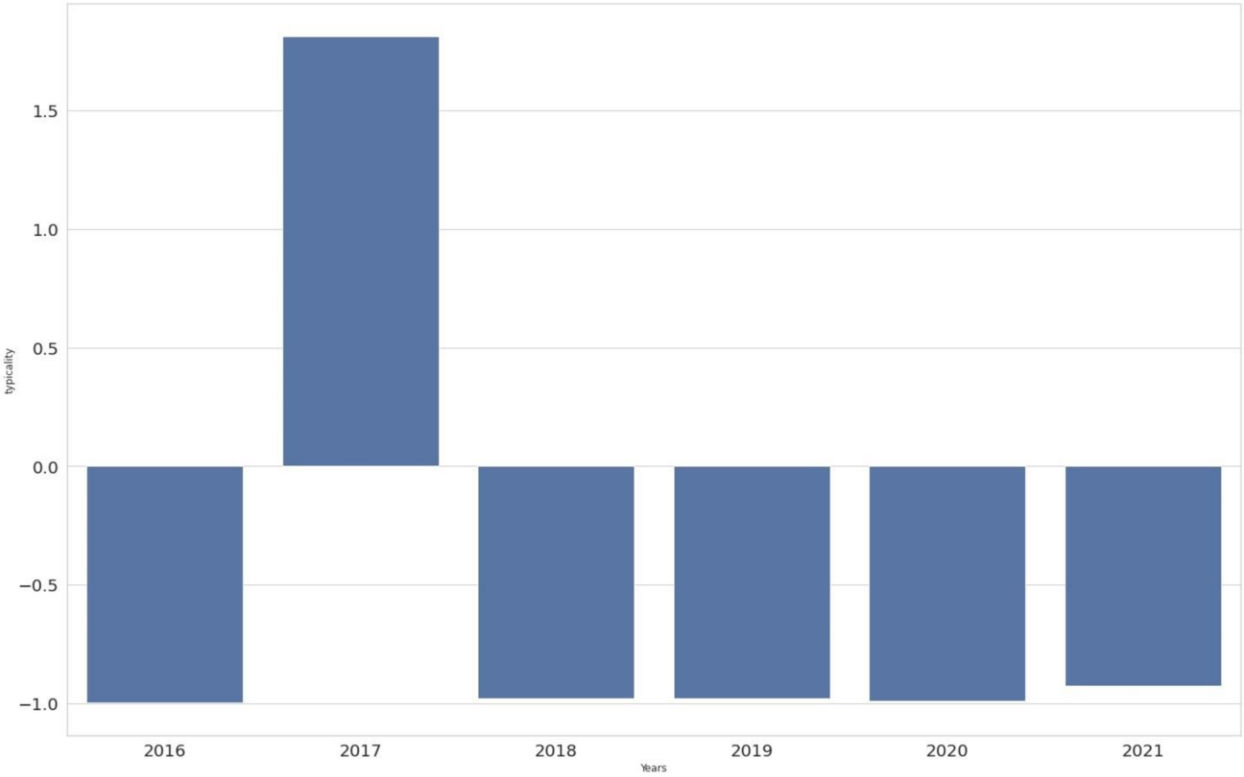


Figure 25 Typicality of the hashtag *muslimban* from 2016 to 2021

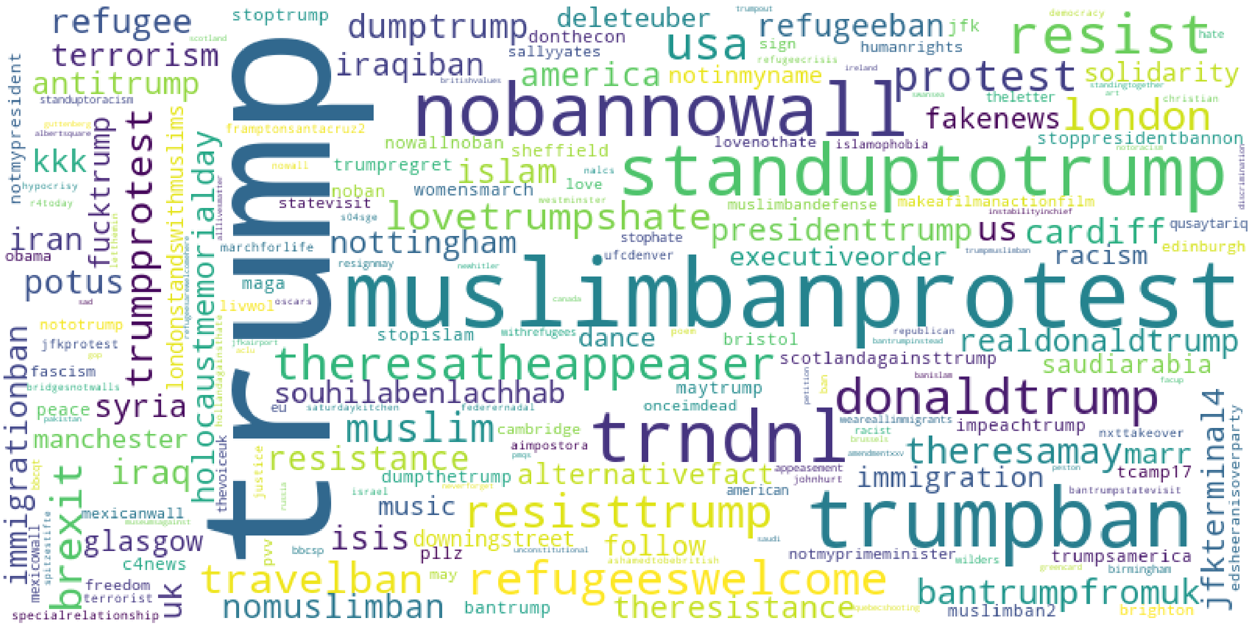


Figure 26 Co-occurring hashtags with *muslimban*

From the co-occurring hashtags, there is more context to be found for the reasons of use of *muslimban.* These hashtags and the peak on the trendline refer to US President Donald Trump's signing of an executive order, preventing Muslims from entering the United States of America (Executive Office of the President 1/27/2017) . Along with this contextual information, the co-occurring hashtags also hint at the opinions expressed by the people on Twitter about this event. With hashtags like *muslimbanprotest, nobannowall* and *standuptotrump* there was a substantial protest and negative sentiments against this executive order.

This approach does illuminate a lot of contextual information also in languages other than English. For example, looking for the hashtags occurring during the June 2018 peak on the German trendline does point towards the event causing this surge of tweets. Figure 28 shows the trendline for German tweets and figure 29 shows the hashtags used.

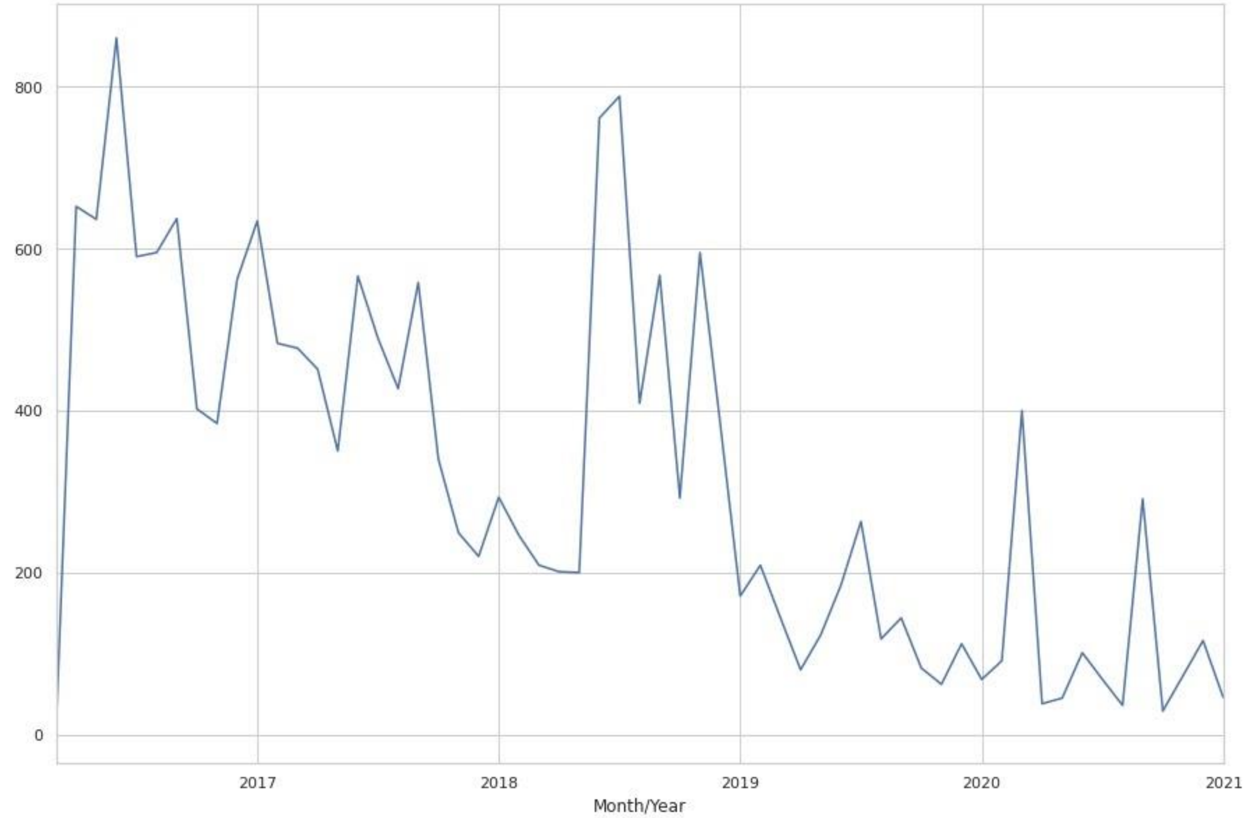


Figure 27 Trendline of German tweets

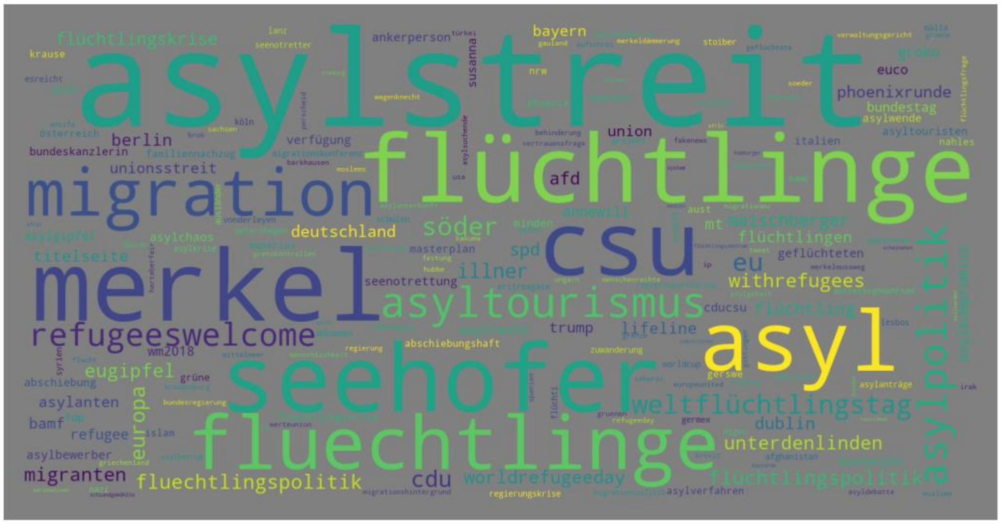


Figure 28 Hashtags used during June 2018 in German

From the figure 29, the hashtag *asylstreit* refers to the political crisis faced by the Merkel government because of the internal dispute between Christlich Demokratische Union and Christlich Sozialen Union (Nielsen 2018) . From the co-occurring hashtags, it is a difficult to properly gauge the opinions expressed during this event. Two hashtags, r*efugeeswelcome* and *withrefugees* does indicate a positive sentiment towards refugees. But it must be noted that *refugeeswelcome* might refer to the project of housing refugees in shared apartments across Europe and Australia (Refugees Welcome International 2021) . Interestingly, the typicality of the hashtag *refugeeswelcome* does coincide with the time of this movement becoming popular across Europe. Most of the European countries began this program of resettling refugees away from camps into their own homes from 2014 till 2016. The hashtag *refugeeswelcome* also is typical during 2016 as seen in figure 30.

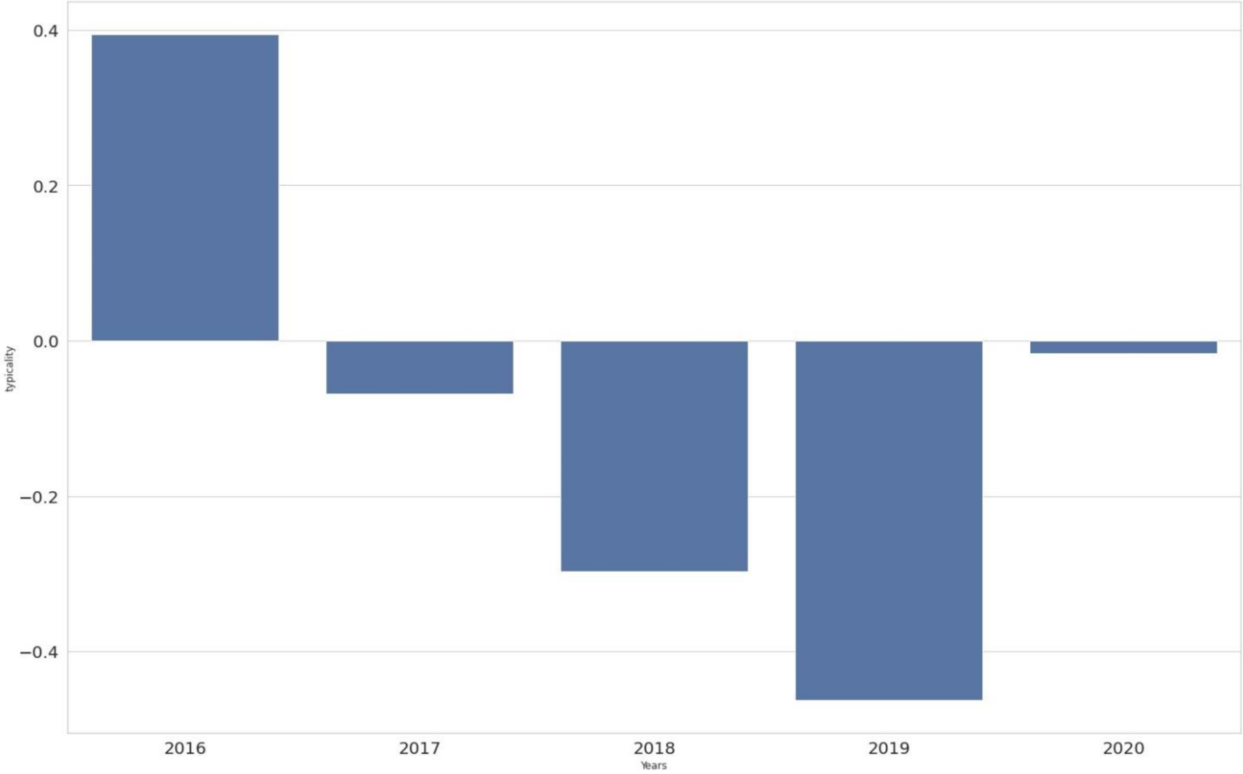


Figure 29 Typicality of the hashtag *refugeeswelcome* from 2016 to 2021

This is not to say that the hashtag *refugeeswelcome* does not indicate positive sentiments towards refugees in general. Afterall the movement does advocate for better living conditions for refugees in the face of insufficient government policy. But the opinion which might be expressed specifically to the government crisis is difficult to determine. The word cloud for the hashtags used during this time along with *refugeeswelcome* are not very indicative of the stance with regards to the government crisis (figure 31). But it does provide further evidence of the pro - refugee stance through two hashtags: *flüchtlingewillkommen* and *keinmenschistillegal.*

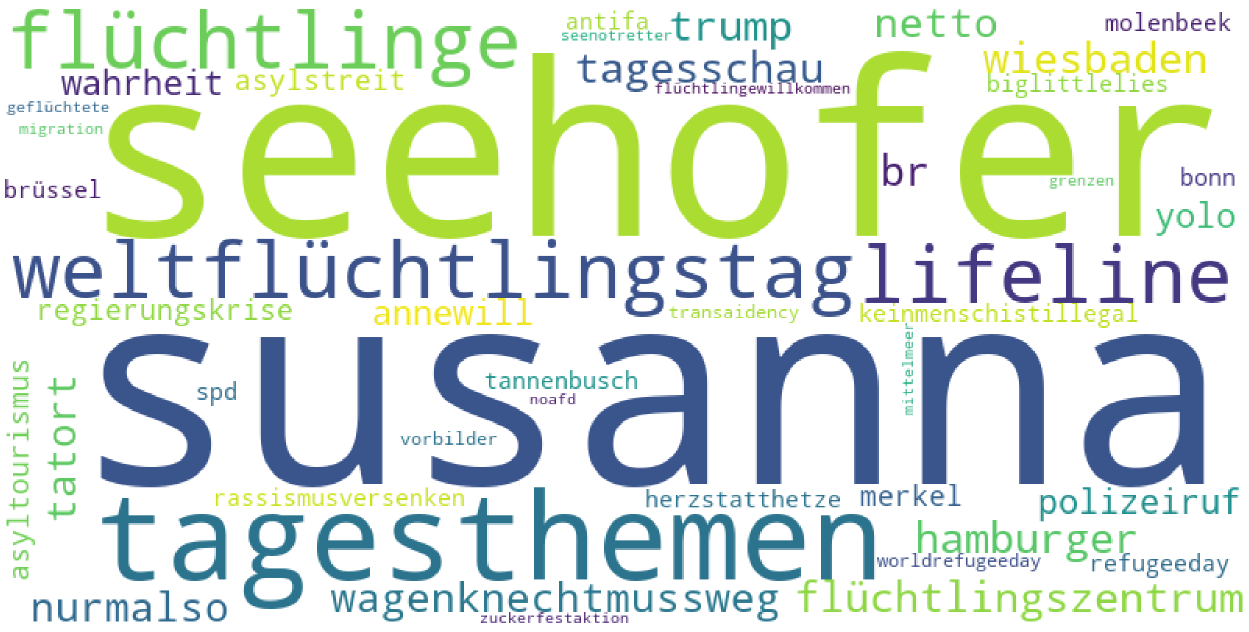


Figure 30 Co-occurring hashtags of the hashtag *refugeeswelcome*

The hashtag *flüchtlingewillkommen* is the German translation of Refugees Welcome, the apartment sharing initiative. *keinmenschistillegal* refers to an interconnected group of activists advocating for rights of migrants (Beckett 2021). The presence of these hashtags does indicate a social grouping and attitude towards migrants and the migration crisis (Laucuka 2018) . This is also true for the anti-migration stance expressed by the hashtag *rapefugees* and its co -occurring hashtags shown in figure 32. *identitarian* refers to a pan- European far right group (Mudde 2019) who had set sail upon a charter boat called Defend Europe (Walsh 2017) . The hashtag *defendeurope* is also seen in the word cloud and could refer to this incident.

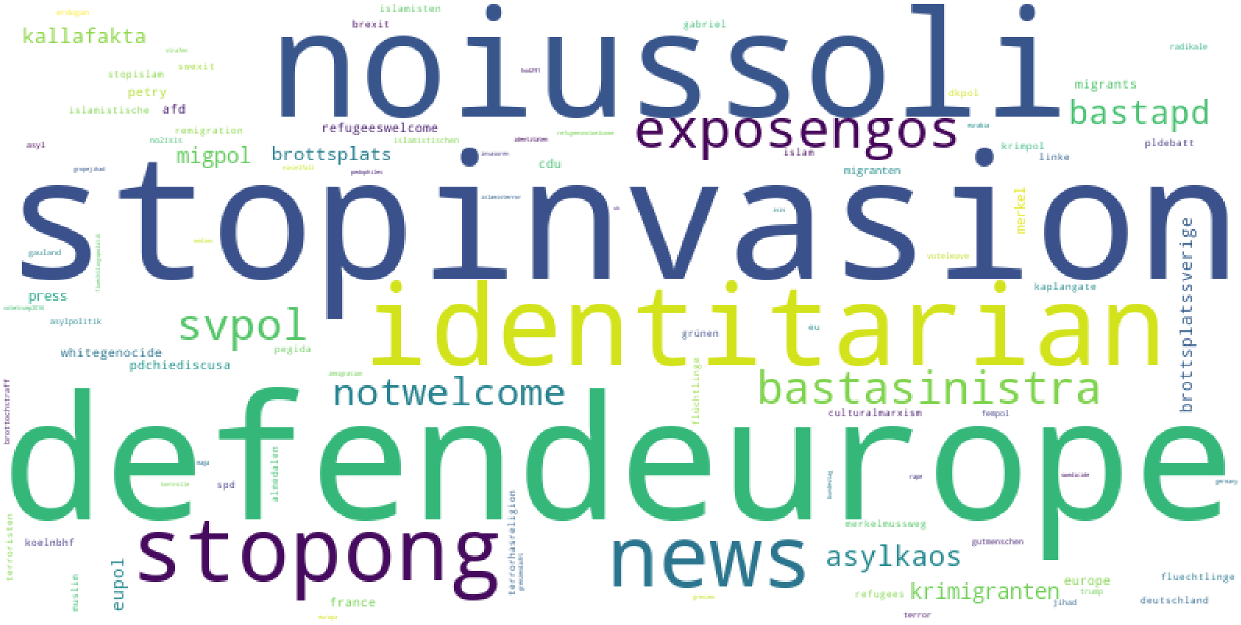


Figure 31 Co-occurring hashtags of the hashtag *rapefugees*

The reasons for very few anti-immigrant hashtags must not be inferred as the overwhelming support for refugees. One hypothesis could be that anti-immigrant sentiments are expressed under hashtags to the political leaders. This is supported by the presence of *rapefugees* as one of the co-occurring hashtags along with *merkel* as seen in figure 33. But this cannot be taken as a definitive conclusion because of the low number of anti-migrant hashtags and a large prevalence of hashtags which cannot be categorized as for or against migrants (neutral hashtags).

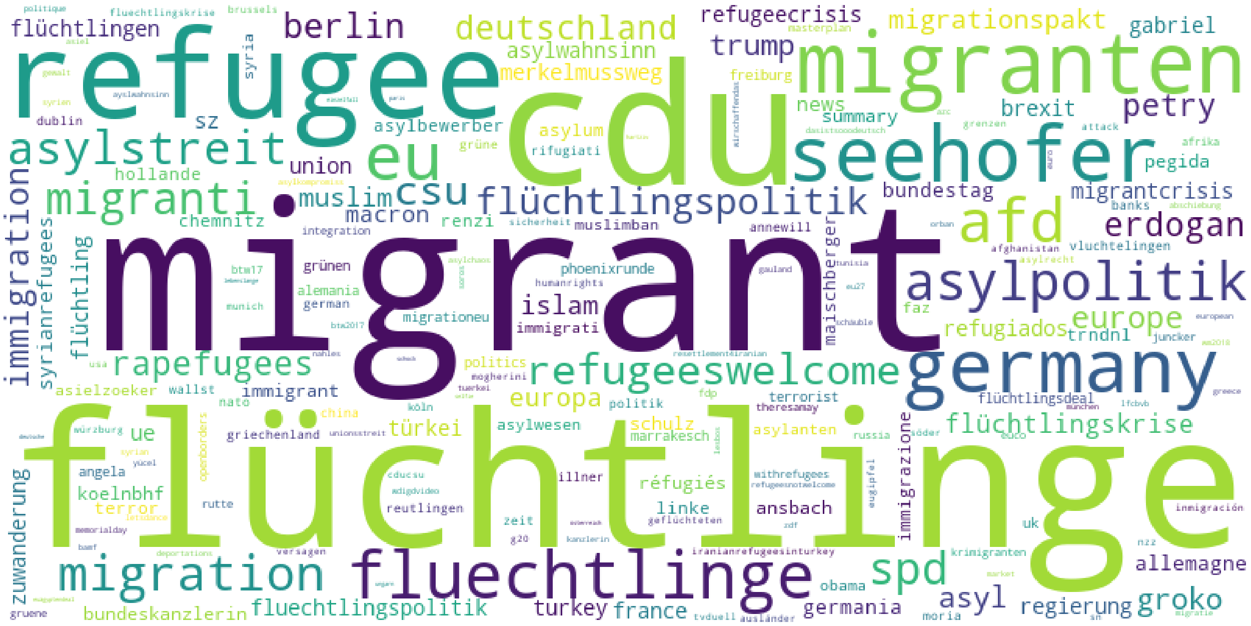


Figure 32 Co-occuring hashtags with *merkel*

Large number of neutral hashtags are also a reason for not being able to definitively assign opinions to events. This is where the author had hoped that sentiment analysis would have been able to add more foundation to the observations previously stated. The multilingual nature of the dataset further caused problems in both event detection and opinion analysis. For German and English, the author could consult authentic citable sources for the events and use them to substantiate the research. The word clouds and events for other languages potentially could support or negate the observations reported above, especially the ones from opinion analysis. For these reasons, the author chooses not to make conclusions out of the mentioned observations.
]

5. Conclusion and Outlook

The ubiquitous use of social media has allowed for researchers to harness a new source of data and apply novel techniques to better understand the numerous social media platforms. This thesis contributed to applying some of these techniques on the case study of the migration crisis in EU. The still ongoing crisis, with it's pan-European nature made this case study uniquely challenging. In spite of the challenges, there are promising results which should be further researched.

Firstly, the privacy aware HLL format does perform as per expectations both in quantitative and qualitative analysis while at the same time protecting the privacy of the users. Even though, the setup for this format could be technically daunting, the lossless union and intersection functions, along with the various metrics allow for accurate observations and analyses. These benefits, unfortunately, come along with certain disadvantages. Primarily among the disadvantages, is the lack of connection between the facets in the HLL format. This prevented the author from fully exploring the dataset using only the HLL format.

Secondly, the dashboard created for visualizing the results of the analyses along with certain elements of narrative visualization makes the dashboard suitable for displaying the multi-faceted twitter data. However, the effectiveness of this dashboard has not been studied and should be a future work. This dashboard falls short of the definition of a dashboard and simultaneously is not a complete narrative visualization. Hence, future work should focus on better integration of the two forms and evaluating its effectiveness.

Lastly, using hashtags for opinion analysis has immense potential in datasets where hashtags are used across various languages. Evidence from this work suggest that both positive and negative opinions can be summarized by looking at hashtags and other co-occurring hashtags. This is at least true, for the English and German tweets. It is also interesting to see that a lot of the popular hashtags has been used and popularized by institutions by both pro-refugee and anti-refugee stance. Throughout this part of the thesis, typicality has proved to be a robust measure for event detection from the hashtags. However, this is not true for every hashtag that has been studied making space for future work in this direction.

Publication bibliography

Aleksey Bilogur; Aneesh Karve; Luis Marsano; Martin Fleischmann (2019): ResidentMario/geoplot 0.3.3: Zenodo.

Alfano, Mark; Reimann, Ritsaart; Quintana, Ignacio; Cheong, Marc; Klein, Colin (2021): The affiliative use of emoji and hashtags in the Black Lives Matter movement: A Twitter case study.

Arnett, George (2015): One in four UK smartphone owners does not make phone calls weekly. In *The Guardian*, 8/9/2015. Available online at https://www.theguardian.com/news/datablog/2015/sep/08/one-in-four-uk-smartphone-weekly-phone-calls, checked on 8/18/2021.

Ayvaz, Serkan; Shiha, Mohammed O. (2017): The Effects of Emoji in Sentiment Analysis. In *(JCEE* 9 (1), pp. 360–369. DOI: 10.17706/ijcee.2017.9.1.360-369.

Beckett, Lois (2021): ‘A system of global apartheid’: author Harsha Walia on why the border crisis is a myth. In *The Guardian*, 4/7/2021. Available online at https://www.theguardian.com/world/2021/apr/07/us-border-immigration-harsha-walia, checked on 8/25/2021.

Blagdon, Jeff (2013): How emoji conquered the world. Available online at https://www.theverge.com/2013/3/4/3966140/how-emoji-conquered-the-world, updated on 3/4/2013, checked on 8/18/2021.

Burghardt, Dirk; Duchêne, Cécile; Mackaness, William (Eds.) (2014): Abstracting Geographic Information in a Data Rich World. Cham: Springer International Publishing.

C. Hutto; Eric Gilbert (2014): VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *ICWSM* 8 (1), pp. 216–225. Available online at https://ojs.aaai.org/index.php/icwsm/article/view/14550.

Caracciolo, Marco (2009): David Herman, Basic Elements of Narrative, Wiley-Blackwell, Oxford 2009. In *Cognitive philology* 2.

Desfontaines, Damien; Lochbihler, Andreas; Basin, David (2018): Cardinality Estimators do not Preserve Privacy. Available online at http://arxiv.org/pdf/1808.05879v3.

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

Dove, Graham; Jones, Sara (2012): Narrative visualization: Sharing insights into complex data.

Dunkel, Alexander; Andrienko, Gennady; Andrienko, Natalia; Burghardt, Dirk; Hauthal, Eva; Purves, Ross (2019): A conceptual framework for studying collective reactions to events in location-based social media. In *International Journal of Geographical Information Science* 33 (4), pp. 780–804. DOI: 10.1080/13658816.2018.1546390.

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

European Union Agency for Fundamental Rights (2020): How concerned are Europeans about their personal data online? Available online at https://fra.europa.eu/en/news/2020/how-concerned-are-europeans-about-their-personal-data-online, updated on 12/14/2020, checked on 8/18/2021.

Executive Office of the President (1/27/2017): Protecting the Nation From Foreign Terrorist Entry Into the United States. E.O. 13769 of Jan 27, 2017. Source: 82 FR 8977, pp. 8977–8982. Available online at https://www.federalregister.gov/documents/2017/02/01/2017-02281/protecting-the-nation-from-foreign-terrorist-entry-into-the-united-states.

Few, Stephen (2006): Information dashboard design: The effective visual communication of data: O'reilly Sebastopol, CA (2).

G. Almatar, Muhammad; Alazmi, Huda S.; Li, Liuqing; Fox, Edward A. (2020): Applying GIS and Text Mining Methods to Twitter Data to Explore the Spatiotemporal Patterns of Topics of Interest in Kuwait. In *IJGI* 9 (12), p. 702. DOI: 10.3390/ijgi9120702.

Hauthal, Eva (2015): Detection, modelling and visualisation of georeferenced emotions from user-generated content.

Hauthal, Eva; Burghardt, Dirk; Dunkel, Alexander (2019): Analyzing and Visualizing Emotional Reactions Expressed by Emojis in Location-Based Social Media. In *IJGI* 8 (3), p. 113. DOI: 10.3390/ijgi8030113.

Hauthal, Eva; Dunkel, Alexander; Burghardt, Dirk (2021): Emojis as Contextual Indicants in Location-Based Social Media Posts. In *ISPRS International Journal of Geo-Information* 10 (6), p. 407. DOI: 10.3390/ijgi10060407.

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

Inuwa-Dutse, Isa; Liptrott, Mark; Korkontzelos, Ioannis (2020): Migration and Refugee Crisis: a Critical Analysis of Online Public Perception. Available online at http://arxiv.org/pdf/2007.09834v1.

Janes A.; Sillitti A.; Succi G. (2013): Effective dashboard design. Cutter IT Journal 26(1). Available online at https://www.researchgate.net/profile/alberto-sillitti/publication/286996830_effective_dashboard_design/links/57c699e208aec24de0414df1/effective-dashboard-design.pdf.

Janvrin, Diane J.; Raschke, Robyn L.; Dilla, William N. (2014): Making sense of complex data using interactive data visualization. In *Journal of Accounting Education* 32 (4), pp. 31–48. DOI: 10.1016/j.jaccedu.2014.09.003.

Jiang, Bin (2013): Head/Tail Breaks: A New Classification Scheme for Data with a Heavy-Tailed Distribution. In *The Professional Geographer* 65 (3), pp. 482–494. DOI: 10.1080/00330124.2012.700499.

Kounadi, Ourania; Resch, Bernd; Petutschnig, Andreas (2018): Privacy Threats and Protection Recommendations for the Use of Geosocial Network Data in Research. In *Social Sciences* 7 (10), p. 191. DOI: 10.3390/socsci7100191.

Kralj Novak, Petra; Smailović, Jasmina; Sluban, Borut; Mozetič, Igor (2015): Sentiment of Emojis. In *PloS one* 10 (12), e0144296. DOI: 10.1371/journal.pone.0144296.

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

Laucuka, Aleksandra (2018): Communicative functions of hashtags. In *Economics and culture* 15 (1), pp. 56–62.

Leskovec, Jure; Rajaraman, Anand; Ullman, Jeffrey David (2020): Mining of Massive Data Sets: Cambridge University Press.

Li, Mengdi; Chng, Eugene; Chong, Alain Yee Loong; See, Simon (2019): An empirical analysis of emoji usage on Twitter. In *IMDS* 119 (8), pp. 1748–1763. DOI: 10.1108/IMDS-01-2019-0001.

Löchner, Marc; Dunkel, Alexander; Burghardt, Dirk (Eds.) (2018): A privacy-aware model to process data from location-based social media.

Löchner, Marc; Dunkel, Alexander; Burghardt, Dirk (2019): Protecting Privacy Using Hyperloglog to Process Data from Location Based Social Networks. In *Proceedings of the Legal Ethical factorS crowdSourced geOgraphic iNformation*.

Madden, Mary (2014): Public Perceptions of Privacy and Security in the Post-Snowden Era. In *Pew Research Center: Internet, Science & Tech*, 12/11/2014. Available online at https://www.pewresearch.org/internet/2014/11/12/public-privacy-perceptions/, checked on 8/18/2021.

Mahajan, Kirti Nilesh; Gokhale, Leena Ajay (2018): Comparative Study of Static and Interactive Visualization Approaches. In *International Journal on Computer Science and Engineering (IJCSE), e-ISSN*, pp. 975–3397.

Mudde, Cas (2019): The Far Right Today: John Wiley & Sons.

Nielsen, Maiken (2018): Fragen und Antworten: Worum geht es bei dem Asylstreit? In *tagesschau.de*, 6/15/2018. Available online at https://www.tagesschau.de/inland/faq-asylstreit-101.html, checked on 8/26/2021.

Öztürk, Nazan; Ayvaz, Serkan (2018): Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. In *Telematics and Informatics* 35 (1), pp. 136–147. DOI: 10.1016/j.tele.2017.10.006.

Parker, Ashley (2011): Hashtags, a New Way for Tweets: Cultural Studies. In *The New York Times*, 10/6/2011. Available online at https://www.nytimes.com/2011/06/12/fashion/hashtags-a-new-way-for-tweets-cultural-studies.html, checked on 8/18/2021.

Pruthi, Purva; Yadav, Anu; Abbasi, Farheen; Toshniwal, Durga (2015): How Has Twitter Changed the Event Discussion Scenario? A Spatio-temporal Diffusion Analysis, pp. 733–736. DOI: 10.1109/BigDataCongress.2015.120.

Refugees Welcome International (2021): Refugees Welcome | This is the international website of the berlin-based project „Flüchtlinge Willkommen". Available online at https://www.refugees-welcome.net/, updated on 2/10/2021, checked on 8/25/2021.

Rosenblatt, Murray (1956): Remarks on Some Nonparametric Estimates of a Density Function. In *Ann. Math. Statist.* 27 (3), pp. 832–837. DOI: 10.1214/aoms/1177728190.

Rubin, Jeff; Chisnell, Dana (2008): How to plan, design, and conduct effective tests. In *Handbook of usability testing* 17 (2), p. 348.

VanderPlas, Jacob; Granger, Brian; Heer, Jeffrey; Moritz, Dominik; Wongsuphasawat, Kanit; Satyanarayan, Arvind et al. (2018): Altair: Interactive Statistical Visualizations for Python. In *JOSS* 3 (32), p. 1057. DOI: 10.21105/joss.01057.

Walsh, Alistair (2017): 'Defend Europe' Identitarians charter a ship to return migrants to Africa | DW | 15.07.2017. Deutsche Welle (www.dw.com). Available online at https://www.dw.com/en/defend-europe-identitarians-charter-a-ship-to-return-migrants-to-africa/a-39702947, updated on 7/15/2017, checked on 8/25/2021.

Wood, Spencer A.; Guerry, Anne D.; Silver, Jessica M.; Lacayo, Martin (2013): Using social media to quantify nature-based tourism and recreation. In *Sci Rep* 3 (1). DOI: 10.1038/srep02976.

Yi, Ji Soo; Kang, Youn Ah; Stasko, John; Jacko, Julie (2007): Toward a deeper understanding of the role of interaction in information visualization. In *IEEE transactions on visualization and computer graphics* 13 (6), pp. 1224–1231. DOI: 10.1109/TVCG.2007.70515.

Yuita Arum Sari; Evy Kamilah Ratnasari; Siti Mutrofin; Agus Zainal Arifin (2014): USER EMOTION IDENTIFICATION IN TWITTER USING SPECIFIC FEATURES: HASHTAG, EMOJI, EMOTICON, AND ADJECTIVE TERM. In *Jurnal Ilmu Komputer dan Informasi* 7 (1), pp. 18–23. DOI: 10.21609/jiki.v7i1.252.

Zeng, Li; Starbird, Kate; Spiro, Emma S. (Eds.) (2016): # unconfirmed: Classifying rumor stance in crisis-related social media messages.

7. Appendix

The dashboard can be found at this link:

https://share.streamlit.io/thecount11/thesis_dashboard/eva_feedback/main.py

It is possible that the dashboard might hang and the maps may take a long time in loading. This is generally fixed by refreshing the webpage. On the off chance, the dashboard might not load at all and display an “Over Capacity” message. In this case, please contact the author at sagnik.rick7@gmail.com and it will be fixed as soon as possible.

The Github repository connected to the dashboard is found under this link. The file *main.py* is the script that runs the dashboard.

https://github.com/TheCount11/Thesis_dashboard

Additional scripts and data which have been used throughout the thesis can be found in this repository:

https://github.com/TheCount11/Carto_Master_Appendix