Information Processing and Management 61 (2024) 103670

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

The voices of the displaced: Mobility and Twitter conversations of migrants of Ukraine in 2022

Richard Lemoine-Rodríguez[a,b,*], Johannes Mast[b], Martin Mühlbauer[b], Nico Mandery[b], Carolin Biewer[a,c], Hannes Taubenböck[a,b,d]

[a] *Geolingual Studies Team, University of Würzburg, Am Hubland, 97074 Würzburg, Germany*
[b] *Earth Observation Center (EOC), German Aerospace Center (DLR), 82234 Oberpfaffenhofen, Germany*
[c] *Department of English and American Studies, Chair of English Linguistics, University of Würzburg, 97074 Würzburg, Germany*
[d] *Institute of Geography and Geology, University of Würzburg, 97074 Würzburg, Germany*

ARTICLE INFO

Keywords:
Twitter
Migration
Topic modeling
Migrant discourses
Ukraine
War

ABSTRACT

Monitoring and understanding human migration as triggered by a crisis is challenging. Combining spatial analysis with natural language processing when analyzing social media data helps to understand the mobility and the needs of migrants better. For this paper, we used geolocated Twitter data to analyze the mobility of and topics discussed by migrants of the Ukraine war in 2022. We removed bots, accounts showing implausible mobility, and automated text content from our dataset. Then, we applied a transformer-based multilingual topic modeling framework to identify the migrants' discourses. We assessed the topics discussed by migrants before leaving Ukraine, after leaving Ukraine and after returning to Ukraine. Our results show that "Attack reports", "politics", "donations to Ukrainians", "food export/production", "humanitarian aid", "nuclear threat", "Ukrainian places", "job search", and "war journalism" were dominant topics before leaving from and after returning to Ukraine. "Food", "social media", "transport", "art", and "finance", however, were important topics right after leaving the country. Overall, our results reveal plausible spatial patterns of migration, which are similar to those reported by official statistics ($R^2 = 0.89$), showing the reliability of geotagged social media data to monitor human mobility. This information can complement official sources, adding first-hand information on the mobility and needs of migrants across space, time, topics, and languages. This is crucial to develop humanitarian response plans when time is of the essence.

1. Introduction

Sadly, human migration triggered by a crisis is an all too common phenomenon these days that poses a number of challenges for monitoring (Chi et al., 2020). Among them, the rapid pace at which migrants move and the scarcity of real-time data make it difficult to quantify migrant mobility (Hübl et al., 2017). Moreover, the process of identifying the migrants' changing needs when moving to new locations over time is time-consuming and resource intensive. To address these challenges, our research focused on Twitter (now called X) data to assess the mobility and needs of migrants who left Ukraine in 2022 after the beginning of the war. With this research being based on Twitter data, we use the term *migrant* in this article in a narrow definition when discussing our results to refer to active Twitter

* Corresponding author: Am Hubland, 97074 Würzburg, Germany.
E-mail address: richard.lemoine-rodriguez@uni-wuerzburg.de (R. Lemoine-Rodríguez).

https://doi.org/10.1016/j.ipm.2024.103670
Received 28 August 2023; Received in revised form 10 January 2024; Accepted 22 January 2024
Available online 30 January 2024

users from Ukraine who, after the start of the war, crossed at least one international border to leave Ukraine temporarily or permanently. First, we assessed the number of migrants per country from February to September 2022, i.e., during the first seven months of the war, controlling for cases of implausible mobility and bots. Secondly, we identified the migrants' needs and interests at different migration stages, using a transformer-based multilingual topic modeling framework. Finally, we validated our findings by comparing the identified mobility patterns and topics with official data from the United Nations High Commissioner for Refugees (UNHCR). Our approach allowed us to identify the needs and interests of migrants across space, time, topics, and languages, opening up new directions for a better understanding of the phenomenon of a migration triggered by a crisis. In the following sections of the introduction, we provide further background on the Ukraine war (Section 1.1), on previous research endeavors addressing this crisis (1.2), and the key contributions of our research (1.3).

1.1. Background

On February 24, 2022, the long-standing conflict between Russia and Ukraine escalated when missile strikes coming from Russia targeted diverse areas of Ukraine (CNN, 2022; IOM, 2022). As of today, almost two years after the beginning of the war, confrontations remain intense in numerous regions within the country, with no signs yet of an end of this ongoing battle (UNHCR, 2023a). Due to this conflict, nearly a third of the Ukrainian population has fled their homes in search for protection, safety, and aid, as part of what has become one of the largest current crises of human displacement in the world (UNHCR, 2023b). As of October 2023, more than 6 million refugees from Ukraine have been recorded around the world, from which ~5 million are currently situated in Europe (UNHCR, 2023c). The European and Asian countries that have received the largest number of migrants from Ukraine since the start of the war are: Poland (~1.5 million), Russia (~1.2 million), Germany (~1 million), the Czech Republic (~0.5 million), Italy (~170,000), Spain (~160,000), the United Kingdom (~160, 000), France (~120,000), Moldova (~100,000), Romania (~100,000), and Slovakia (~100,000; UNHCR, 2023c). This mass migration is symptomatic for a humanitarian crisis, which has been characterized by a set of challenges when it comes to the provision of fundamental needs to the refugees – from ensuring their safety, and addressing legal issues, to providing them with long-term support, and managing the public perception in the host countries (UNHCR, 2023b).

To cope with these challenges, the UNHCR and other partner institutions have been conducting surveys in Poland, Bulgaria, Slovakia, Moldova, Romania, and Hungary to collect data on the social background and needs of refugees from Ukraine (UNHCR, 2022a, 2023d). Interviews have been carried out at border crossings, reception centers, collective sites, and assistance centers in major cities since May 2022 to date (i.e., October 2023). From these, personal data such as nationality, gender, age, and education level, as well as the new place of residence and the urgent needs of migrants in the host countries have been identified. The results indicate that material assistance, food, employment, health care, accommodation, education, transportation, and financial aid are the most critical needs for the migrants from Ukraine (UNHCR, 2022a, 2023d).

Since the second half of 2022, a high number of migrants has been recorded as returning to Ukraine. It is estimated that >1 million people returned to the country despite warnings from Ukrainian officials about the still hostile and dangerous situation (IOM, 2022). While some people returned to Ukraine to stay, others went back to pick up relatives, personal documents, and material belongings to take them out of the country, resulting in multiple entries to and exits from Ukraine. For those migrants who eventually returned to Ukraine, financial support, heating appliances, medicine, sanitary products, health services, food, and clothes have been found to be among the most urgent needs (IOM, 2022).

In a migration crisis, it is essential for politicians, decisions-makers, and stakeholders alike that information on paths and patterns of mobility be rapidly transmitted to allow for a timely and helpful response plan. Data on human migration is commonly recorded in strategic border crossing areas and administrative centers by means of surveys and censuses, as well as in local offices in the host countries, e.g., when asylum seekers are registered (Chi et al., 2020; UNHCR, 2023d). These official sources can provide valuable, consistent, and systematic information on the migration flows, which often is publicly available. Nevertheless, a complex international standardized survey strategy and coordination must be implemented to conduct successful campaigns to record the refugees, and developing such a strategy may delay the start of the data collection (Mazzoli et al., 2020). Moreover, these surveys are based on samples taken at a certain point in time and are not representative of all stages of the migration crisis. Furthermore, once the information is collected, it may take some time until it is processed and made publicly available. This in turn may delay the process of optimizing mobility routes and transport, and preparing host countries to provide sufficient adequate accommodation and first aid (UNHCR, 2023b). It is necessary to identify the specific needs of migrants at different host locations over different stages of migration to implement adequate humanitarian response plans throughout the migrants' journey, and that as quickly as possible (Hübl et al., 2017).

1.2. Related work

The increasing employment of "citizens as sensors" or "social sensing" has proved to be a game-changing approach for the monitoring of human migration (Chi et al., 2020; Hübl et al., 2017; Li et al., 2021; Mast et al., 2023). The usefulness of Twitter data for monitoring human migration has been validated through a comparative analysis of geolocated Tweets and mobile phone trace data, revealing that, based on Twitter data, residents can be distinguished from non-residents over extended timeframes (Chi et al., 2020). In the context of a crisis, geolocated Twitter data was used in 2015 to study the migration of refugees from the Middle East and North Africa to Europe; the study included the quantification of Tweets referring to refugees along the migration routes based on keywords and hashtags (Hübl et al., 2017). Due to an increasing interest in monitoring human mobility through Twitter data, there has been a recent development of scalable online platforms which enable the extraction, the analysis, and the visualization of origin-destination flows (Li et al., 2021). These platforms aim to facilitate the assessment of Twitter data without requiring special expertise in quantitative data analysis.

Social media datasets reveal the topics users are interested in and their opinions, offering to researchers the possibility to assess, among others, the discourses constructed by migrants (Khatua & Nejdl, 2021). Thanks to its free accessibility for research up to the beginning of 2023, Twitter data has been increasingly employed to either identify human mobility patterns or assess discourses referring to refugees on social media, and, in few cases, to combine both approaches (Arcila-Calderón, et al., 2022; Chi et al., 2020; Hübl et al., 2017; Khatua & Nejdl, 2021; Mast et al., 2023; Smith et al., 2018). Twitter represents a valuable data source to cope with

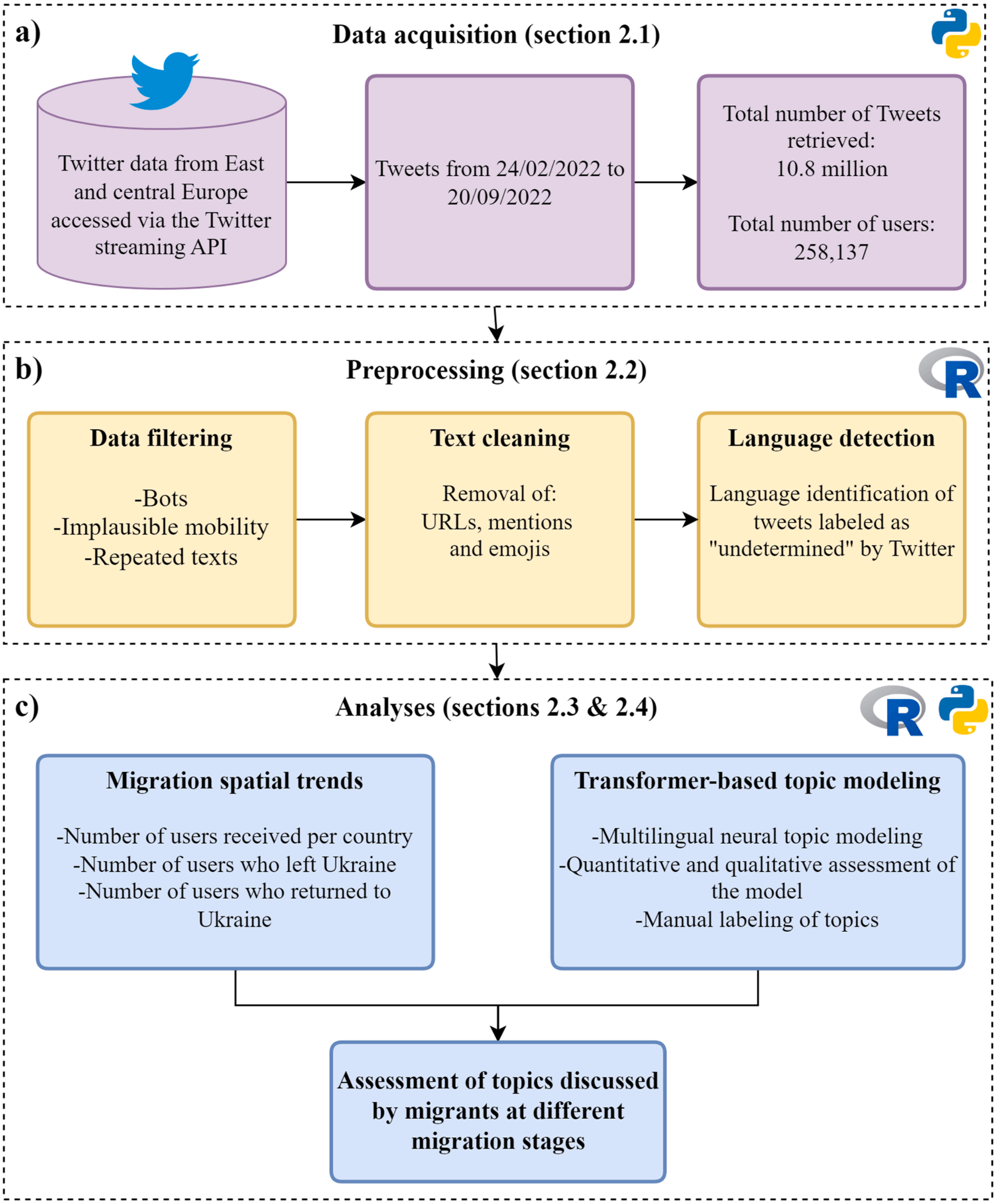

a)
Data acquisition (section 2.1)
Twitter data from East and central Europe accessed via the Twitter streaming API
Tweets from 24/02/2022 to 20/09/2022
Total number of Tweets retrieved: 10.8 million
Total number of users: 258,137
b)
Preprocessing (section 2.2)
Data filtering
-Bots
-Implausible mobility
-Repeated texts
Text cleaning
Removal of: URLs, mentions and emojis
Language detection
Language identification of tweets labeled as "undetermined" by Twitter
c)
Analyses (sections 2.3 & 2.4)
Migration spatial trends
-Number of users received per country
-Number of users who left Ukraine
-Number of users who returned to Ukraine
Transformer-based topic modeling
-Multilingual neural topic modeling
-Quantitative and qualitative assessment of the model
-Manual labeling of topics
Assessment of topics discussed by migrants at different migration stages

Fig. 1. Steps conducted to assess the mobility of and topics discussed by the Twitter users who left Ukraine after the start of the war concerning: (a) data acquisition, (b) preprocessing stages, and (c) types of analysis.

these tasks in hand, since it is widely used around the world and it includes information such as user information, the texts the users posted, the date of content creation, and often their geolocation (Senaratne et al., 2023). This fulfills the requirements of digital trace data (i.e., digital data including location and time stamps), enabling researchers to assess the users' mobility through spatial analysis (Chi et al., 2020). Moreover, the content of the texts shared by Twitter users constitutes a basis to identify the needs and interests of the users, as well as their opinions and emotions, with the help of natural language processing (NLP) techniques (Grootendorst, 2022; Pota et al., 2021) and linguistic know-how (Martin & White, 2005).

In response to the war in Ukraine, diverse research groups and organizations from around the world have been using social media and spatial data to analyze, evaluate, and monitor the impacts of this humanitarian crisis. Integrating digital traces from Facebook and census data, the displacement of people inside Ukraine has been mapped with geo-spatial methods (Leasure et al., 2022b). The authors estimated population sizes and changes on a daily basis, aggregated by age and sex for each primary administrative unit of the country. Based on reports from Twitter, Telegram, Facebook, and news articles, the presence of remaining explosives in Ukraine was mapped, as well as building and infrastructure damage, employing geographical information systems (GIS; Lanclos & Cottray, 2022). With the help of geospatial technologies, the damage to diverse types of infrastructure was also identified. Nighttime aerial imagery was used, and a 26% decrease in light emission from Kyiv was measured after a wave of missile and drone strikes in the city (Conflict Observatory, 2023). Multiple efforts have been carried out to implement artificial intelligence systems to automate damage detection based on satellite images and other datasets (Lanclos & Cottray, 2022; Shamoug, 2022). Moreover, NLP was applied to sift through text reports and find ways to identify the parties involved automatically, together with the time and the location of the strikes, as well as the type of damage to the infrastructure (Shamoug, 2022). This approach was complemented by a sentiment analysis of Twitter data, to assess the moods and the needs of people residing in affected areas (Vahdat-Nejad et al., 2023).

Using social media data to develop a crisis response plan faces several challenges. Social media data is not representative of the entire population, due to its user base usually being mainly composed of young to middle-aged individuals from urban areas (Zhu et al., 2022). In terms of content, data such as Twitter posts are difficult to interpret, since they are composed of short texts which often only contain a few words, emojis, pictures, URL's or hashtags (Pota et al., 2021). This complexity increases by the presence of automated Twitter accounts (i.e., bots), making it challenging to assess the mobility and opinions of migrants reliably as they relocate during a crisis (Rodríguez-Ruiz et al., 2020; Schuchard et al., 2019).

1.3. Contribution

As a contribution to this collective effort of responding sensibly to the humanitarian crisis triggered by the war in Ukraine, we integrated spatial analysis and NLP techniques in this research to identify the mobility patterns and topics discussed by Twitter users who left the country. We analyzed Tweets posted from the start of the war until September 2022. We implemented filters to exclude bots, and to control the reliability of the identified mobility and the quality of the text content. We followed the trajectories of Twitter users leaving and, in some cases, eventually returning to Ukraine, and learned about the main topics they discussed before leaving, after leaving and after returning to the country (Fig. 1). This approach allowed us to identify the interests and needs of the migrants at specific stages of their migration employing first-hand multilingual information from social media in real-time. This approach and the gained information may (1) serve as a rapidly produced indicator of mobility trends and the needs of migrants in future crises, (2) complement humanitarian response plans, adding the migrants' own voices to the official information and surveys, and (3) help to understand and respond better to the migrants' diverse interests and needs in different hosting locations as they continue their journey.

2. Methods

2.1. Data acquisition

We aimed to assess the mobility and topics discussed by Twitter users who left Ukraine after the beginning of the war. For that purpose, we acquired geolocated Twitter data from Eastern and part of central Europe posted between February 24, 2022, and September 20, 2022 (~7 months; Fig. 1a). The query area was defined by a bounding box extending from 43.16, 56.70 North to 11.78, 42.95 West (cf. Fig. 5). The data was accessed through the Twitter streaming application programming interface (API) employing the Tweepy python library (Roesslein, 2020). All Tweets included metadata information related to the users (e.g., user id, user description, number of followers, number of accounts followed by the corresponding user, date when the account was created) and metadata related to each Tweet (e.g., Tweet id, Tweet text, date and time of creation, country, language, place, place type, place coordinates, source, and in some cases, precise coordinates). The two coordinate types included in the Twitter metadata correspond to (1) GPS coordinates of the place where the Tweet was posted, and (2) place location, i.e., coordinates linked to a "Twitter place", which may correspond to different "place types" (i.e., granularity levels) such as point of interest (POI), neighborhood, city, admin (i.e., administrative), or country.

2.2. Preprocessing

Datasets collected from Twitter are not analysis-ready due to, among other issues, the presence of bots, and the potential manipulation of geolocations. Thus, we developed and applied various preprocessing steps, i.e., bot filters (cf. Section 2.2.1), mobility filters (cf. Section 2.2.2) and text filters (cf. Section 2.2.3) to exclude automated accounts, users with implausible mobility, and non-relevant Tweets in terms of their text content (i.e., repeated text predefined by third-party apps). Next, we computationally identified the language of the Tweets that had an “undetermined language” label according to their metadata information (Fig. 1b). All pre-processing steps were conducted on the environment for statistical analysis R (R Core Team, 2020).

2.2.1. Bot filter

Due to the high presence of bots on the Twitter platform, it was necessary to apply bot detection techniques to exclude them from our analysis focused on human users (Orabi et al., 2020; Rodríguez-Ruiz et al., 2020). The objectives of bot accounts may range from advertising products or services, informing on specific events (e.g., sports events, natural disasters), to reporting the weather forecast, and sharing news, among others. Beyond this, some also have malicious purposes such as spreading fake news. Multiple methods have been developed to detect Twitter bots. To identify potential bots, most approaches quantify the number of followers and followed accounts, the account activity (e.g., number of posts per day), the interaction with other users, or assess the polarity of the Tweets (Orabi et al., 2020; Rodríguez-Ruiz et al., 2020). In this research, we developed an empirical approach based on the user description – in some cases, automated accounts are openly revealed as such, the ratio between followed and following accounts, and the number of posts over time. We did not exclude accounts based on pre-established thresholds for the estimated indicators, but empirically identified outliers (accounts with uncommon activity) based on our data. Several automated Twitter accounts declared themselves as bots in their user description. We employed this information to remove all the explicitly declared bot accounts from our database (Fig. 1b). To identify non-declared bots, we computed a normalized follower-to-following ratio, based on the fact that bots tend to have more followers than the users they follow (Twitter, 2022). The range of the normalized follower-to-following ratio is −1 to 1. A value of 1 indicates a high number of followers and a low number of followed accounts. In contrast to that, −1 indicates a low number of followers and a high number of followed users. The assumption here is that accounts with high values are potentially bots. After testing the filtering of manually identified bots based on different threshold values, we removed users with a normalized follower-to-following ratio >0.7 who followed <10 accounts.

In addition, we excluded users who posted identical texts more than once or more than one Tweet at the same time from different locations, as this behavior suggests bot activity. We explored the histograms of the distribution of the total number of Tweets posted per day and the average number of Tweets per day for each user, to empirically find threshold values to identify uncommon behavior. Based on this, we removed users who tweeted >50 times in any single day and users with >10 Tweets on average per day.

2.2.2. Mobility filter

The information in the metadata on the geolocation of the Tweets, in theory, would allow us to map the mobility of Twitter users. Nevertheless, since this information can be manipulated by the users, it was necessary to develop and apply filters to exclude implausible mobility (Fig. 1b). Since we aimed to assess migrants from Ukraine, we only included Twitter users whose first Tweet was posted in Ukraine and who left the country during our period of study.

Twitter users may post multiple times from the same place and geotag their Tweets to different locations. To address this issue, we computed the speed of movement between every pair of geotagged Tweets posted by each user in chronological order, based on the Euclidean distance between their coordinates and the time difference between Tweets. Then, from the statistical distribution of the speed of movement between each pair of coordinates by user, we identified outliers. Considering this and the transport possibilities (i.e., car, bus, train, plane), we excluded users with >20 % of their Tweets with a speed of movement >150 km/h, assuming that migrants would not travel at a higher speed several times.

Another aspect to consider regarding the mobility of migrants is the number of times that they leave and return to their country of origin. After exploring the distribution of the number of exits from and returns to Ukraine of each migrant, we excluded users who left or returned to the country >3 times within the seven months of our period of analysis, since a higher number of such trajectories was uncommon according to the data and implausible in a war scenario.

2.2.3. Text filter

It is common to find predefined automated content on social media, produced by the platforms, apps or services from which users repost content. This is variable and depends on the social network as well as the regional user preferences. Since our objective was to analyze original content, we systematically examined the text of the Tweets of our database in the different languages, to identify repeated and predefined text content (Fig. 1b). Based on this analysis, we removed Tweets containing the same text pattern from a single or from multiple users. Most of these Tweets were re-postings from other platforms such as Instagram, containing text in the different languages of our database stating that the users had “just published a photo” or “video”. As we focused on plain text, we also removed the URLs, mentions and emojis from each Tweet.

In Twitter data, as far as their metadata language information is concerned, it is common to find a considerable number of posts labeled as undetermined (i.e., tagged with the “und” code). These Tweets may contain phrases or words in >1 language, may be composed mainly of URLs or emojis, or may contain nothing but symbols that cannot be recognized as a known human language by algorithms. Aiming to include Tweets with undetermined language coding which could be relevant to our analysis, we extracted these Tweets from our database and removed their hashtags (Fig. 1b). After this, we employed two different language identification algorithms to assign a language to them. We applied Google’s Compact Language Detector 2 (CLD2), which probabilistically identifies over 80 languages (Ooms, 2022a). For mixed-language input, CLD2 returns the top main languages found and their approximate percentages for the text in total (e.g., 75 % English and 25 % Spanish). Additionally, we also employed Google’s Compact Language Detector 3 (CLD3; Ooms, 2022b). CLD3 is a neural network model for language identification and the successor of CLD2. We manually verified the accuracy of the results and then retained only Tweets for which the same language label was suggested by both algorithms, since we found that this was the most reliable approach. Here, a single label was added to multilingual Tweets only if ∼90 % of their

Table 1
Volume reduction after applying the filters to exclude bots, implausible mobility, and repeated and automated text content.

	Number of Tweets	Number of users
Total acquired	10,805,122	258,137
After only including users whose first Tweet was in Ukraine and posted more than once	993,325	14,772
After bot filter	846,820	11,309
After mobility filter	82,374	1464
After text filter	54,815	1182

text was written in a single language (Fig. 1b). After this, we reintegrated the Tweets with their new language label into our main database. Sections 2.3 and 2.4 below describe the subsequent data analysis steps used to assess the mobility of the Twitter users and the topics discussed by them (Fig. 1c).

2.3. *Mobility analysis*

The mobility of Twitter users was assessed based on the total data volume after applying the mobility filter (i.e., before applying the text filter). Here, we used the GPS and the place coordinates that were included in the metadata of the Tweets (Fig. 1c). When both coordinate types were available for one Tweet, we employed only the GPS coordinates. Since most of our data (i.e., ~50,000 Tweets) only contained place coordinates at the country level granularity, we will discuss our mobility assessment based on this scale.

We then validated our results by comparing the number of migrants detected from the Twitter data with the official refugee records by country produced by the UNHCR in the last month of our assessment (i.e., September 2022; UNHCR, 2022b). We applied a linear regression to evaluate the validity of our data sample and compare it to the official records only for countries which were fully covered by our bounding box.

We were able to identify mobility from Ukraine to other countries, between countries other than Ukraine, and from other countries to Ukraine. Based on the country where Tweets were located and their chronological order per user, we classified the Tweets into their corresponding migration stages, i.e., as posted before leaving, after leaving, and after returning to Ukraine. For Twitter users with >1 movement corresponding to the same migration stage, we added a label indicating its occurrence number in chronological order (e.g., for a user who left Ukraine three times: after leave for the first, for the second, and for the third time).

2.4. *Topic modeling*

After applying the bot filter, the mobility filter, and the text filter, we applied topic modeling to the remaining Tweets to identify the topics discussed by the identified migrants (Fig. 1c). In this step, we assessed only languages including >1500 Tweets, to have a representative data volume.

We employed BERTopic, a transformer-based neural topic modeling framework that has proven to outperform traditional NLP models (Grootendorst, 2022). Unlike other topic modeling techniques such as Latent Dirichlet Allocation (LDA) or Structural Topic Models (STM) that employ the bag-of-word representations, neglecting word order and context, BERTopic contextualizes word meanings based on its preceding and subsequent words in a sentence (Devlin et al., 2019; Grootendorst, 2022). Due to the use of language models pre-trained on a large text corpus, this produces more interpretable and context-aware topic clusters (Egger & Yu, 2022).

First, the text of the Tweets was converted to sentence embeddings, which are numerical representations of texts. Here, texts with similar meaning are represented with close values, showing their semantic relationship. To do such a conversion, we employed the transformer-based BERT paraphrase-multilingual-mpnet-base-v2 language model, obtained from the HuggingFace Hub (Reimers & Gurevych, 2019; https://huggingface.co/). This model has shown state-of-the-art performance to align multiple languages and it is trained for >50 languages. The dimensions of the word embeddings were reduced employing the Uniform Manifold Approximation and Projection (UMAP) technique, which preserves more local and global features of highly dimensional data than the principal component analysis (PCA; McInnes et al., 2018). We employed the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm to cluster the reduced embeddings into topics (Grootendorst, 2022). HDBSCAN allowed us to group all derived outliers in a single cluster, which avoids forcing them to be part of specific topics in contrast to other algorithms (e.g., K-means). Finally, the class-based term frequency-inverse document frequency matrix (c-TF-IDF) was calculated for each topic. From it, the key terms that best represented each cluster were extracted to help in the thematic interpretation of potential topics.

A priori, the number of topics was unknown. Thus, we tested the performance of the clustering procedure quantitatively and qualitatively. We based this on iterations from 20–100 topics, defining a minimum cluster size of 100 Tweets. We computed the normalized pointwise mutual information (NPMI) of our final model as a quantitative coherence metric (Bouma, 2009). NPMI has been shown to emulate human judgment with reasonable performance (Lau et al., 2014). A highly incoherent model would have an NPMI value of −1, whereas an extremely coherent set of topics would be 1. Nevertheless, our final number of topics was mainly based on the capability of our model to represent diverse general and war-related topics which were interpretable and coherent.

For each topic, we extracted a random sample of 50 Tweets per language (or the maximum available). These Tweets were machine translated into English, and their thematic coherence was assessed independently by two scholars. The assessment of the topics was undertaken based on their key terms from the c-TF-IDF and their text content. Topics which lacked thematic coherence and were

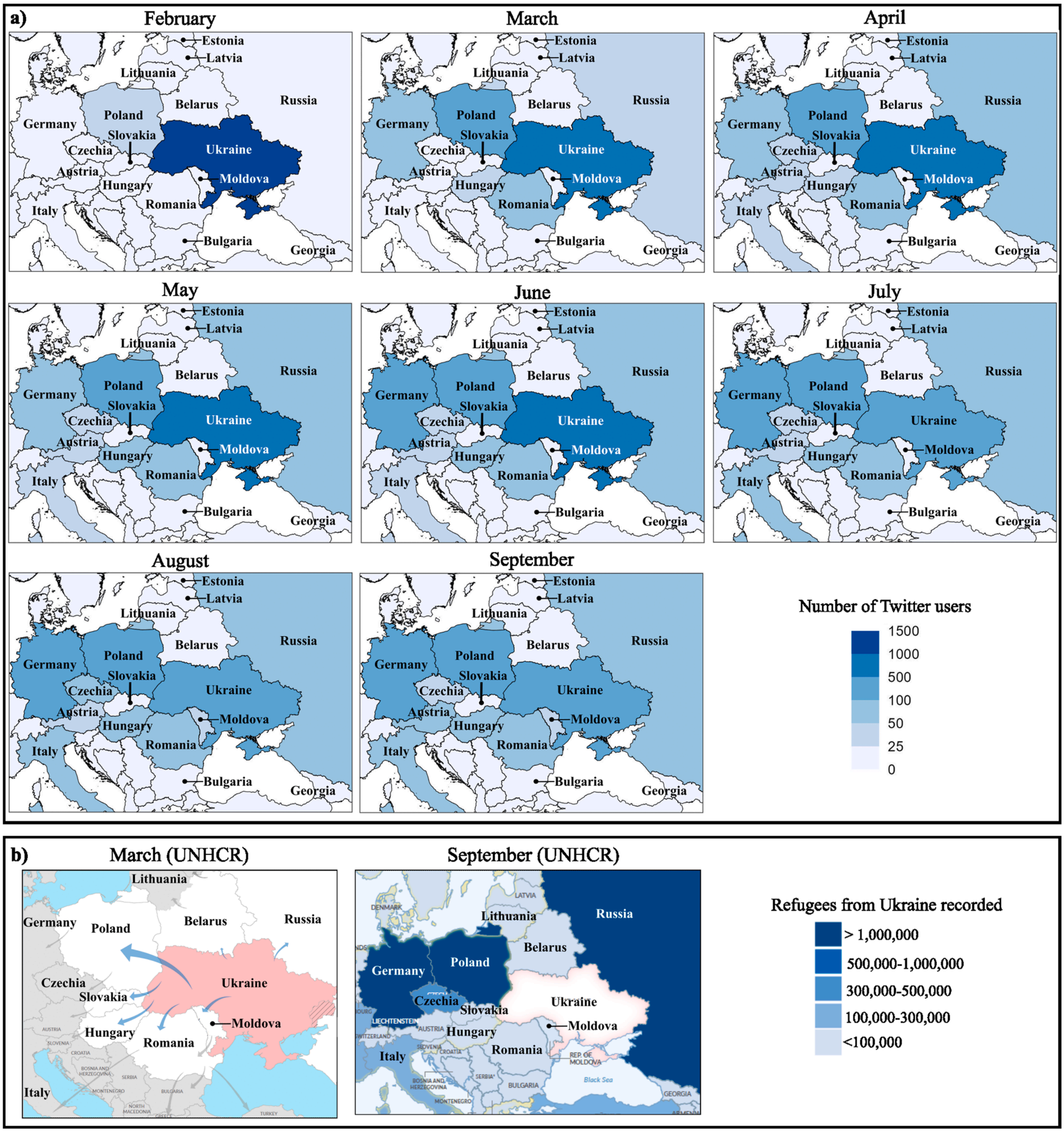


Fig. 2. Monthly migrants from Ukraine by country: (a) as identified from Twitter data, and (b) as of March 29 (left) and September 20, 2022 (right) reported by the UNHCR (UNHCR, 2022b, 2022c). Panel (a) shows the number of migrants by country based on the location of the last Tweet posted by each user in the corresponding month. Panel (b) left (i.e., March) shows the main routes of migrants leaving Ukraine. The size of the arrows represents the proportion of movements compared to the total registered by the UNHCR. Panel (b) right (i.e., September) depicts the total amount of refugees by country recorded by the UNHCR as of September 20, 2022.

therefore too heterogenous for an unambiguous interpretation were integrated into the group of outliers, whereas highly similar topics were merged. A final label was assigned to each topic on which the two researchers agreed. We then validated the identified topics by comparing them with the primary needs of Ukrainian refugees in European countries, as determined through survey campaigns conducted by the UNHCR, 2022a, 2023d, 2023e). Finally, the presence and frequency of the topics was assessed for each migration stage (i.e., before leaving, after leaving, and after returning to Ukraine; Fig. 1c).

3. Results

3.1. Data volume after filtering

In our study area covering Eastern and parts of Central Europe (cf. the bounding box in Fig. 5), we acquired ~10.8 million Tweets posted by 258,137 users (Table 1). The number of Tweets and users in our database was then reduced due to the implemented bot filter, mobility filter, and text filter. The largest data reduction occurred when including only users whose first Tweet was posted in Ukraine and who had posted more than once, resulting into 993,325 Tweets posted by 14,772 users (Table 1). From the filters applied to exclude bots, implausible mobility and repeated and automated text content, the largest volume reduction occurred after applying the mobility filter, reducing the number of Tweets and users ~10 times compared to the totals after the bot filter, decreasing from 846,820 to 82,374 Tweets and from 11,309 to 1464 users. Finally, after the implementation of the final text filter, the total data volume was reduced to 54,815 Tweets posted by 1182 users who left Ukraine during our period of study (Table 1).

3.2. The mobility patterns and Twitter activity of the migrants

The per-country presence of Twitter users who left Ukraine shows the temporal evolution of the migrants' mobility during our period of study (Fig. 2a). In February 2022, at the beginning of the war, most Twitter users who left Ukraine first moved to the neighboring country of Poland. During March, other neighboring countries such as Russia, Romania and Hungary showed an increasing presence of newly arriving Twitter users from Ukraine. Simultaneously, Germany started to emerge as a relevant host country at this early stage of the war. In April, Austria, the Czech Republic, and Italy also exhibited an increasing presence of newly arriving migrants. In the following months the number of migrants in the aforementioned countries continued to rise, as the number of Twitter users in Ukraine decreased. As of September 2022 (i.e., the last month of our period of study), the countries with the highest presence of Twitter users from Ukraine were Germany, Poland, and Ukraine itself. This shows the prevalence and intensification of the main migration route towards Germany through Poland, which first developed in March. The other main countries which exhibited the presence of migrants from Ukraine in September were Russia, Italy, Austria, Hungary, and Romania, followed by the Czech Republic and Moldova (Fig. 2a).

The comparison of our results with the official records shows high consistency. Despite the different orders of magnitude of the data, the number of migrants identified from Twitter data shows a highly linear relationship with the official statistics from the UNHCR as of September 2022, with a correlation of R^2=0.89 (Appendix A). From the countries who were completely covered by our bounding box, Bosnia and Herzegovina and Poland are the ones which deviate the most from a linear pattern compared to the UNHCR records.

After the application of the mobility filter, we identified a total of 1464 Twitter users who left Ukraine during our study period. From them, around one third (i.e., 529) returned to Ukraine within the first 7 months of the war. According to the Twitter data, these migrants showed a total of 5889 movements, from which 1712 were local (i.e., inside Ukraine), 1734 to leave Ukraine, 772 international movements (i.e., they were moving through other countries than Ukraine), 1022 inside the same foreign country (other than Ukraine), and 649 to return to Ukraine. The migration from and to Ukraine shows very dynamic flows, with most of them aiming to leave the country (Fig. 3). The largest peak of movements to leave Ukraine occurred within the first days of the war (i.e., right on and after February 24, 2022). Other peaks of movements to leave the country occurred at the end of March, the beginning of May and July, mid-July, and around August 22 and September 5, 2022 (Fig. 3). In contrast, the highest number of returns to the country took place at the end of March, around May 9, and at the beginning of July (Fig. 3).

According to the first Tweet posted by each migrant outside of Ukraine, most of the 1734 movements to leave the country were in the direction of Poland (656), followed by Russia (181), Germany (176), Romania (130), Hungary (94), the Czech Republic (80), and Italy (69; Fig. 4). These countries were also the ones from which most migrants returned to Ukraine according to their Twitter activity. From the 649 movements aiming to return to the country, 268 Tweets were geotagged in Poland, 94 in Russia, 57 in Germany, 53 in Romania, 34 in Hungary, 29 in the Czech Republic, and 13 in Italy before the migrants returned to Ukraine. The countries from our database which received the least number of migrants right after leaving Ukraine were Sweden (8), Bosnia and Herzegovina (5), Latvia (4), Kosovo (1), and San Marino (1). From them, Twitter users located in Bosnia and Herzegovina, Latvia, and San Marino did not exhibit a subsequent direct movement to return to Ukraine (Fig. 4).

Most movements aiming to leave Ukraine reveal the capital of or other large cities within the foreign countries as the first destination (Fig. 5). From the countries neighboring Ukraine, the cities receiving the highest number of migrants were Moscow in Russia, Bucharest in Romania, Budapest in Hungary, Prague in the Czech Republic, Poznan and Warsaw in Poland, and Vilnius in Lithuania. Other cities with a high number of arrivals were Berlin in Germany, Copenhagen in Denmark, Vienna in Austria, Venice in Italy, and Belgrade in Serbia. In contrast to the trajectories for the movement to foreign countries, most returns to Ukraine were geotagged at country granularity level by the Twitter users (Fig. 5). Due to this, several flow lines representing returns to the country as illustrated in Fig. 5 show the center of the polygon of Ukraine as destination. Despite of this, based on the Tweets geolocated at a more detailed scale (e.g., city, neighborhood), the city of Kyiv appears as one of the main first destinations after returning to Ukraine. The cities with the highest number of migrants returning to Ukraine were Moscow in Russia, Warsaw in Poland, Berlin in Germany, Prague in the Czech Republic, Vienna in Austria, and Budapest in Hungary (Fig. 5).

The number of Tweets posted by the migrants per migration stage naturally decreases, as this number is linked to a reduced number of Twitter users experiencing all different stages in their migration process (Fig. 6). From the total of 82,374 remaining Tweets after applying the bot filter and mobility filter, 55,797 were posted before leaving, 25,928 after leaving, and 649 after returning to Ukraine. In particular, the 1464 identified migrants showed the highest activity on Twitter before leaving Ukraine for the first time, posting

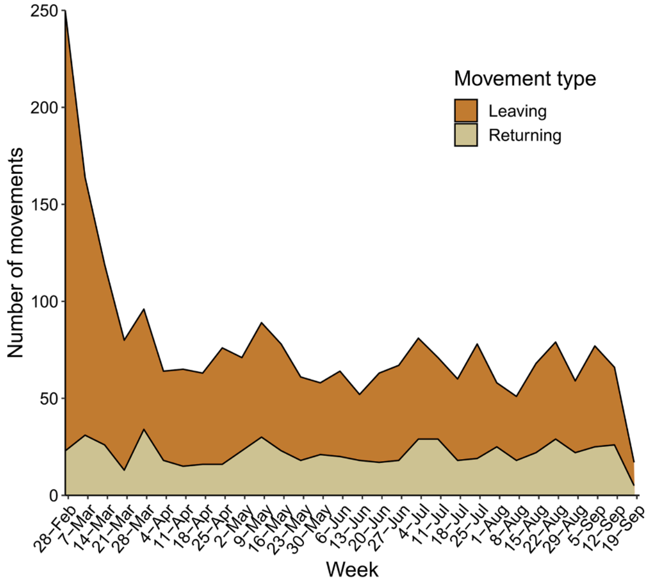


Fig. 3. Number of movements to leave from/to return to Ukraine during our period of study.

35,913 Tweets at that stage (Fig. 6). The number of Tweets was reduced to 22,112 after the first time they left the country. From them, 529 users (i.e., almost one third) returned to Ukraine, posting 18,012 Tweets. 215 migrants exited from the country for a second time, posting 3521 Tweets at that stage. Nearly half of these users (i.e., 103) returned to Ukraine for a second time, producing 2335 Twitter posts. The 55 migrants who left the country for a third time posted a total of 295 Tweets. Finally, 186 Tweets were posted by the 17 migrants who exhibited a third return to Ukraine (Fig. 6).

Within our Twitter dataset, 20 % of the users (i.e., 293) produced 80 % of all the Tweets, following the Pareto principle and showing a power distribution (Appendix B, panel a). Upon exploration through a log-log representation, the frequency of Tweets, ranked by their respective counts per user, shows a pattern approximating Zipf's distribution. However, there is a notable deviation towards the upper end of the plot, departing from a linear shape (slope = −1.35; Appendix B, panel b). This shows that the number of Tweets produced by the users in our database is not inversely proportional to their ranks.

3.3. Topics discussed by migrants at different migration stages

From the total number of Tweets after applying the text filter (i.e., 54,815), most were posted in the Ukrainian language (24,476), followed by English (20,945), Russian (7770), and Polish (1624). After quantitatively and qualitatively assessing the outcome of different amounts of topics from 20–100, a final set of 31 topics with an NPMI value of 0.012 and a high human interpretability confirmed by the two above-mentioned scholars was chosen. Due to the presence of several outlier Tweets (i.e., Tweets which were not semantically similar to others and/or non-interpretable), our final model comprised of 8700 labeled Tweets. The topics in our model represent general themes as well as war-related issues, and they reflect patterns in the interests of the migrants that are related to their migration stage (and therefore their location; in or out of Ukraine; Fig. 7; Appendix C).

Overall, "Attack reports" was the topic most frequently addressed by the migrants in our database. A decreasing number of Tweets about attacks in Ukraine was recorded once migrants were situated out of Ukraine; the Tweets increased again once the migrants returned to the country. "Politics", "donations to Ukrainians", "food export/production", "humanitarian aid", "nuclear threat", "Ukrainian places", "job search", and "war journalism" were also more widely discussed as long as the users were in Ukraine, either before leaving or after returning to the country. Topics such as "art", "transport", and "photography" were frequently addressed after leaving and after returning to the country. "International support/NATO", "languages", and "Nazism/fascism" were frequent topics before and after the migrants left Ukraine, but not after they returned. "Foreigners leaving Ukraine" was not frequently discussed by people who returned to Ukraine, this topic showing its main relevance only before and after leaving the country for Twitter users. "Love" was mainly present in Tweets posted after users left Ukraine. The topics with similar relevance at all three migration stages were: "transport", "literature", "photography", and "vegetation" (Fig. 7; Appendix C).

Regarding the number of Tweets per topic across the final set of languages of our database (i.e., English, Ukrainian, Russian, and Polish), English is the only language represented in all topics (Fig. 8; Appendix D). "Donations to Ukrainians", "international support/NATO", "Russian government", "foreigners leaving Ukraine", "nuclear threat", "humanitarian aid", and "war journalism" were almost exclusively discussed in English. The main topics discussed in the Russian language were: "attack reports", "food", "social media", "finance", "pets", "movies/videos", and "Nazism/fascism". Most Tweets in the Polish language referred to "attack reports", "food", and "transport". The topics which were dominant in Tweets in English and Ukrainian were: "attack reports", "transport", "politics", "languages", "job search", and "encouraging Ukraine". Topics such as "food", "social media", "finance", "pets", "literature", "religion",

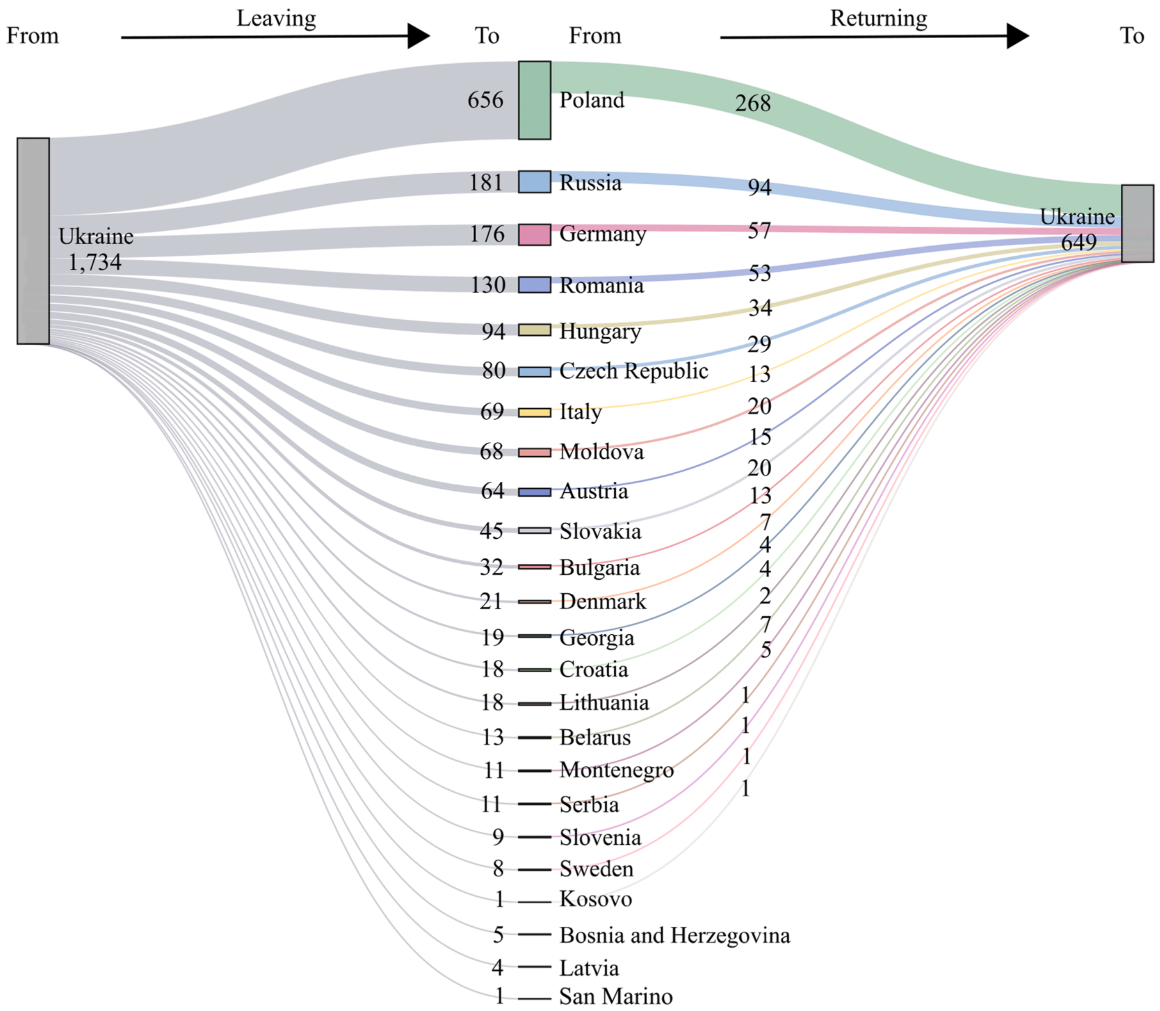


Fig. 4. Number of movements from Ukraine to the first country from which Twitter users posted after leaving the country (left to center) and from the last country before returning to Ukraine (center to right).

"health care", "clothes", and "vegetation" were mainly discussed in Ukrainian (Fig. 8; Appendix D).

Collected under the identified topic labels one finds Tweets which report war statistics, personal situations, personal needs, opinions, and emotions at different migration stages (see examples in Table 2). Topics which refer to everyday issues, such as "pets", "languages", "health care", "job search" or "transport", may actually refer to these issues in relation to the dynamics of the war. Some of the Tweets contain political statements on the impact of the war (e.g., the "languages" topic example in Table 2), others contain descriptions of how pets adapt to the war situation (e.g., under "pets"), they depict desperate attempts of Ukrainians to leave the conflict zones (e.g., under "transport"), the difficulties of some to gain access to health care (e.g., under "health care"), and the struggle of others to find a new income source in this crisis (e.g., under "job search"). Other Tweets which are directly connected in content to the war situation show casualty statistics, even related to particular groups such as journalists (e.g., under "war journalism"), and attack reports and good-bye messages when a possible fatal outcome of the next attack is anticipated by the user (e.g., under "attack reports"). The call for international support to protect Ukraine or provide armory to the country was also present in multiple topic clusters (e.g., in "int. support/NATO"). The fear of facing a new nuclear disaster if Russia might use nuclear weapons or attack a nuclear power plant framed a whole topic cluster on its own in our model (i.e., "nuclear threat"). Despite all these challenges and negative experiences described by the users, migrants also exhibited positive feelings (e.g., when they marvel at their dog's cleverness to understand what hearing a siren means) or sent messages of support and encouragement to other Ukrainians to keep fighting (e.g., "encouraging Ukraine").

4. Discussion

In response to the ongoing war in Ukraine, in this research we used geolocated Twitter data to identify mobility patterns and the communicated needs of migrants who left the country within the first seven months of the war. Based on our results, we will discuss the

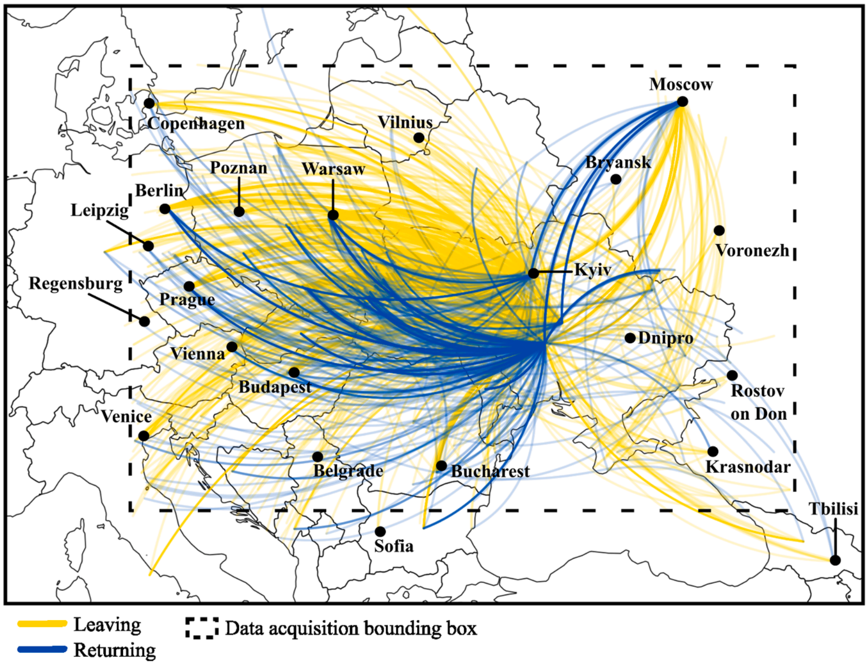


Fig. 5. First destination of migrants after leaving from (yellow) and last destination before returning to Ukraine (blue). Several movements are represented by precise, POI, city, or admin coordinates. Nevertheless, it can be noticed that for Germany, Ukraine, Bulgaria, Montenegro, Georgia and Italy, multiple Tweets are located at the center of the corresponding country bounding box. This is due to Tweets geotagged at the country level.

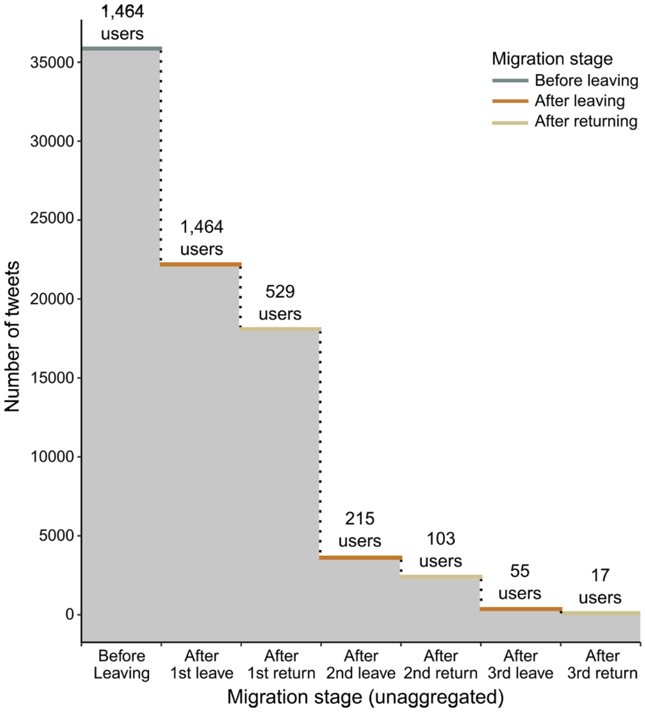


Fig. 6. Number of Tweets posted at each unaggregated migration stage and its corresponding number of Twitter users.

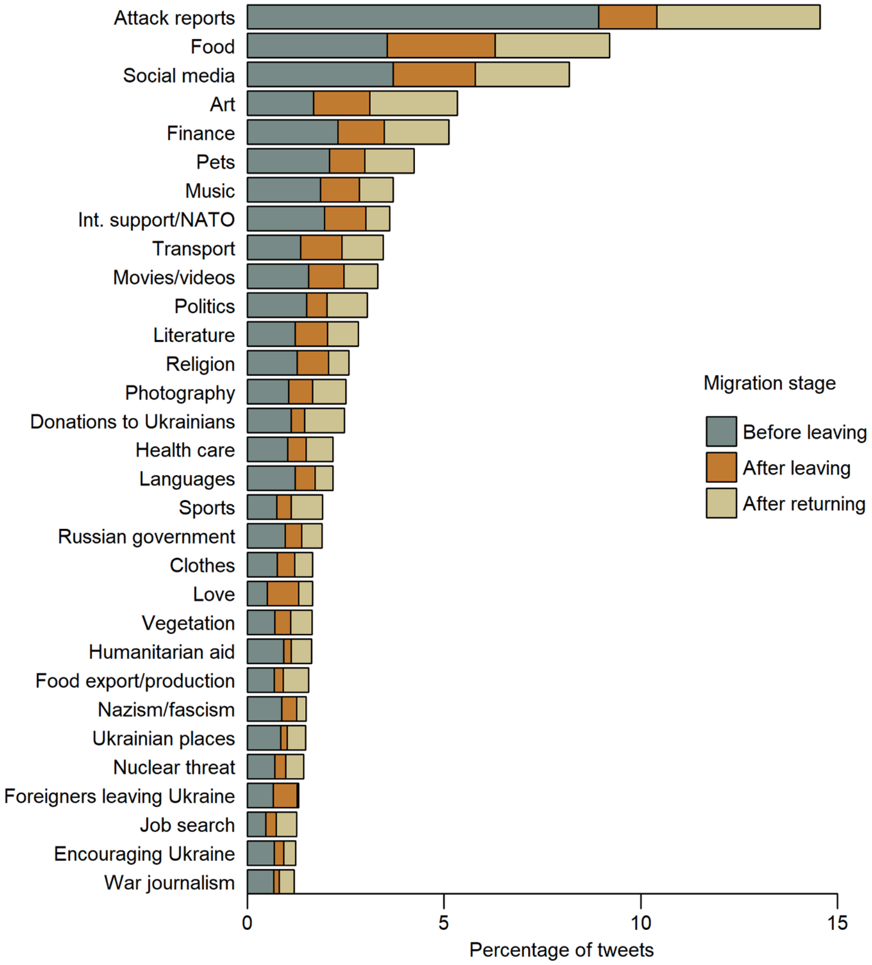


Fig. 7. Percentage of Tweets related to each topic by migration stage.

following issues in this section: (1) the spatial migration patterns identified with the help of the Twitter data in comparison to official statistics, (2) the main topics discussed by migrants at different migration stages compared to the needs of migrants from Ukraine as identified by traditional surveys, and (3) the challenges of analyzing social media datasets and how they were addressed in this study.

4.1. Migration patterns: Twitter data versus official statistics

Compared to the official number of migrants who fled the war in Ukraine recorded by the UNHCR, our sample shows a small proportion of the people who left the country. Nevertheless, when looking at the per country presence of migrants and their paths of movement at different stages during the war (e.g., at the beginning and seven months later), our results and the official statistics strongly correspond (cf. Fig. 2). At an early stage of the war (i.e., at the end of March 2022), Poland was identified by the UNHCR to be the main destination outside Ukraine, with subsequent movements to Germany and the Czech Republic (UNHCR, 2022c). The neighboring countries of Slovakia, Hungary, Romania, and Moldova, followed by Russia, were also identified by the UNHCR as countries with a relevant amount of migration flows. Overall, we found the same patterns (although on a much smaller scale) based on Twitter data (cf. Fig. 2). Compared to the official records, the main difference is that Slovakia, Moldova, and the Czech Republic did not show a high presence of migrants from Ukraine by the end of March in our data. In the last month of our period of study (i.e., September), we identified Germany and Poland as the countries with the highest presence of migrants from Ukraine. As of September 20, 2022, the same countries plus Russia were officially recorded by the UNHCR as the main hosts receiving people from Ukraine (UNHCR, 2022b). A possible reason why our data seems to indicate that Russia is less relevant than Germany and Poland as a host country may be the lack of accessibility to several internet sites and platforms such as Twitter in Russia. Other countries which received a high amount of migrants as of September 2022 according to the UNHCR are the Czech Republic and Italy (UNHCR, 2022b). Our data also shows Italy at the same level of relevance as Austria, Hungary, and Romania as host country of migrants from the war by the end of September – even though the bounding box we used to acquire the data only covered a small portion of Italy (cf. Fig. 5). In our study, we mainly focused on international migration due to ~50,000 Tweets geolocated at the country level (i.e., with not much of an

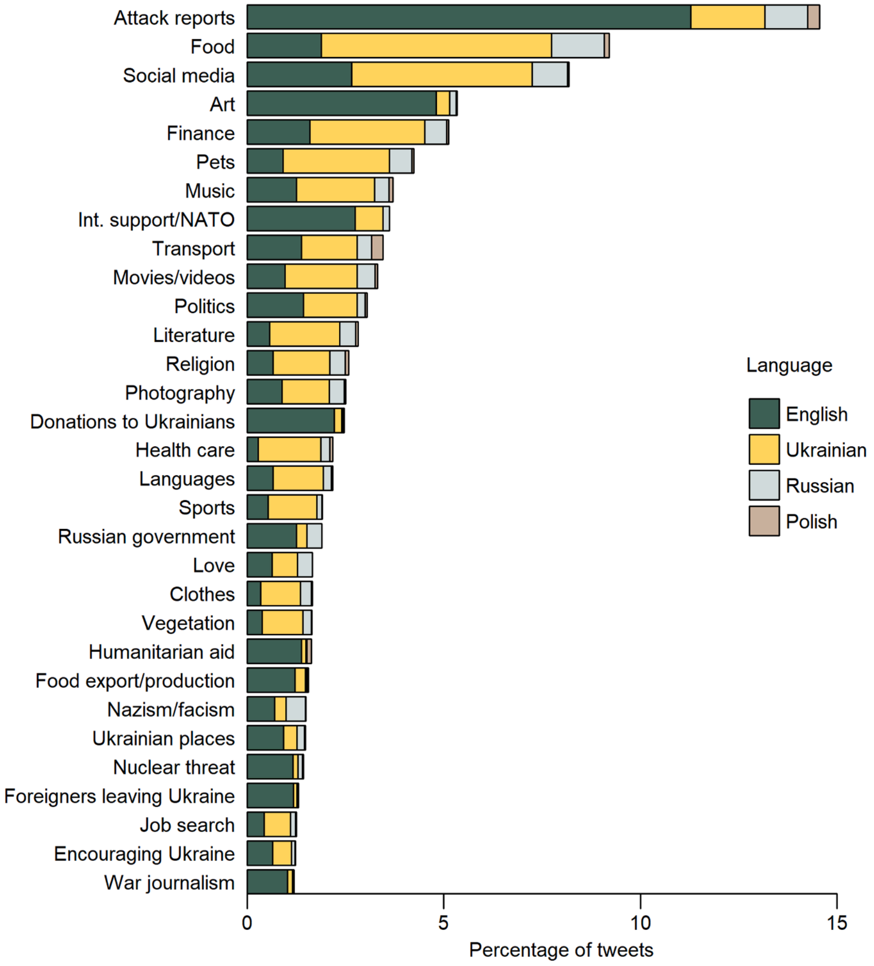


Fig. 8. Percentage of Tweets related to each topic by language

indication where exactly within the country the Tweets were posted). As a consequence, our research did not include the assessment of mobility within Ukraine, where the war produced intensive internal migration (UNHCR, 2023b). If the volume of Tweets geo-referenced below the country level is higher, intra-national migration analysis is also possible following the same approach as Mast et al. (2023) have shown for Nigeria. As domestic mobility as a result of a humanitarian crisis can be substantial, it is important to also monitor this phenomenon, if possible, to help developing targeted strategies to address the needs of the people affected and fostering some form of recovery and resilience within the nation.

As acknowledged above, we used a more vague and more restricted definition of *migrant* for this study since we used this label for active Twitter users from Ukraine who, after the start of the war, crossed at least one international border to leave Ukraine temporarily or permanently. Within the dataset we did not further differentiate between various reasons for such international mobility within the first seven months of the war. The migrants that we identified, therefore, potentially include Twitter users who were travelling to Russia or other countries without the intention to migrate. In this research, a validation of the Twitter users who were residents (as in Chi et al., 2020) in Ukraine before the war was not possible since we acquired data only from the start of the conflict. Despite these limitations, our results derived from Twitter data show mobility patterns which are consistent with the official statistics in terms of the rank of countries and cities receiving the highest number of migrants from Ukraine. Moreover, the highly linear relationship between the migrants that we identified by country at the end of our period of study (i.e., September 2022) and the official statistics from the UNHCR shows the plausibility of our approach (cf. Appendix A). It can therefore be concluded that in this case of an imminent crisis triggering mass migration in 2022 Twitter was a platform that showed reliable statistics on the movement of people leaving Ukraine.

Social media datasets from platforms such as Twitter have allowed to identify specific cities to which migrants arrived during other crises (Hübl et al., 2017; Mazzoli et al., 2020). Our results show that the main destinations of migrants were capitals and other large cities in the host countries. This may be motivated by the media popularity of capital cities and the fact that large cities have a higher capacity to provide access to jobs, education, health, and other social and technical infrastructure compared to small ones (Bettencourt et al., 2007; Elmqvist et al., 2021). Additionally, it is possible that migrants who moved to small and remote settlements experienced poorer internet access. Therefore, those individuals may be underrepresented in the Twitter data. Another defining characteristic of the

Table 2
Examples of Tweets posted by migrants after the start of the war.

Text	Topic	Migration stage	Original language
"Journalists covering #UkraineRussiaWar have paid a high price. The total killed so far represents about 20 % of all journalists killed worldwide in 2021."	War journalism	After returning	English
"I am proud of Ukraine, my Motherland. I am proud of every Ukrainian who does his best. who makes cocktails, who unarmedly stops a column of tanks, who informs about the movements of the occupiers. I am proud that I am Ukrainian Everything will be fine, Ukraine!"	Encouraging Ukraine	Before leaving	Ukrainian
"there are more than five rocket hits near me, the nearest 3 min on foot from me, if something happens to me, know that I loved you"	Attack reports	Before leaving	Ukrainian
"I love how my dog is hearing the siren and starts pushing me to get out of the bed and go to our bomb shelter each time. Smart cookie!"	Pets	After returning	English
"The change in language use in the last six months has been palpable even in #Odesa. When I came here in March, virtually no one I talked to spoke Ukrainian. Now an ever-increasing number of people, especially young ones, make it a point to not speak Russian anymore. #Ukraine"	Languages	After returning	English
"I'll distract you, otherwise no one knows - my back is broken in two places, the spinal cord is transferred. This creates pain levels of 12/10, which I could not even guess. In this state, I was taken out of Kyiv and I have been living on wild droppers / injections / painkillers for a week now"	Health care	Before leaving	Russian
"It's like this in Ternopil now: an electric train from Kharkov has just arrived. Several thousand passengers left. We rode standing, sitting, in the vestibule. Anyhow. 23 h. I'm trying to leave for Kyiv, but the trains are delayed. The most incredible thing is that now all trains run for free. This is what they told me in the manual"	Transport	Before leaving	Russian
"NATO, its a need #NoFlyZone! If not, equip Ukraine to protect our own sky. …stop bombs and missiles, save lives of people #protectUAsky #airdefenceforUkraone #jetsforUkraine"	Int. support/ NATO	Before leaving	English
"The threat level at the Zaporozhye nuclear power plant is very high, the Ministry of Internal Affairs is preparing for any scenario, up to the evacuation of the population, said the head of the Interior Ministry"	Nuclear threat	After leaving	Polish
"Life" is so interesting.. one day you have two jobs, another day zero Let's start this "fascinatingly" journey: how to find a job during the war. Take a popcorn, it's going to be interesting #DEVCommunity #job #vacancies #javascript"	Job search	After returning	English

selected migrant destinations during the Ukrainian migration crisis was social connectedness (Best & Menkhoff, 2022). Among other things, it was found that regions with Facebook users who are digitally connected to users from Ukraine showed a higher number of migrants from Ukraine. This seems to point to the migrants' preference to arrive at places where they already know people, which may enhance opportunities to find accommodation and jobs (Best & Menkhoff, 2022).

When migrants arrived at a new host country, many of them were officially registered as asylum seekers, so it was possible to record the number of migrants in different countries to a certain extent (UNHCR, 2023b). Contrastingly, the locations from where migrants returned to their country of regular residence (Ukraine in this case) are less clear in the official statistics since there is no motivation for them to inform the local authorities of the (temporal) host country about their departure. Digital trace data from social media may help in this regard, allowing to find the locations from which social media users are digitally active before returning to their country of origin (Chi et al., 2020). We identified Poland, Russia, Germany, Romania, and Hungary as the main countries from which migrants posted on Twitter for the first time after leaving Ukraine. This is in line with the official findings in terms of the main exit routes from Ukraine (UNHCR, 2022b, 2022c). Additionally, we identified that these same countries were the ones from which the Twitter users posted for the last time before returning to the Ukraine. The relevance of Poland, Romania, Hungary, and Russia as host countries can be explained by their geographical proximity, being neighboring countries of Ukraine. Moreover, in the case of Russia and Poland, its cultural, historical, and linguistic ties with Ukraine might be an important factor attracting migrants fleeing the war, many of them with relatives living in these countries.

4.2. The migrants' interests and needs reflected by Twitter and survey data

The topics that we identified allowed us to document opinions, emotions, concerns and needs of migrants from Ukraine. These varied depending on their migration stage, since they faced different scenarios before leaving their country, after leaving, and after returning (UNHCR, 2022a, 2023b). NLP techniques allowed us to classify the migrants' stories and discussions across languages and therefore trace similarities and differences in the choice of topics in the different languages of our sample. In this sense, the methodological contribution of our approach lies in the use of a transformer-based multilingual topic modeling framework to assess the discourses of migrants, which better captures the semantic context of Tweets compared to previous word-based topic classifications (Hübl et al., 2017).

We found that Twitter users mainly discussed "attack reports", "politics", "donations to Ukrainians", "food export/production", "humanitarian aid", "nuclear threat", "Ukrainian places", "job search", and "war journalism" when they were in Ukraine (i.e., before leaving and after returning to the country). Twitter users informed other users about dangerous places and situations, the need for international aid, and the concerns about food supplies, income sources and the risks faced by journalists covering the war. "International support/NATO", "languages", and "Nazism/fascism" were relevant issues before and after migrants left Ukraine. This shows the need for support from the international community and the issues related to communication that some migrants faced (i.e., the use of a specific language as a political statement or communication barriers in foreign countries). Additionally, several Tweets were used to express political opinions. While migrants stayed outside Ukraine, topics such as "food", "social media", "art", "finance", "transport", "music", "international support/NATO", and "attack reports" were frequent in the Twitter conversations. At this stage, Twitter users were still sharing information on attacks in Ukraine and requesting international support. However, other topics on the economic situation and on leisure activities were gaining in importance. A key finding when looking at the migrants' needs as expressed on Twitter is the language in which they communicate specific issues. Topics such as "donations to Ukrainians", "international support/NATO", "Russian government", "foreigners leaving Ukraine", "nuclear threat", "humanitarian aid", and "war journalism" were almost exclusively discussed in English. This shows the attempt to reach an international audience when calling for aid (i.e., "donations to Ukrainians", "international support/NATO", "humanitarian aid"), when sharing concerns over risks with far-reaching consequences (i.e., "nuclear threat"), and political opinions (i.e., "Russian government") as well as personal experiences as foreigners in Ukraine (i.e., "foreigners leaving Ukraine").

The main needs of the migrants identified by ground surveys (UNHCR, 2022a, 2023d, 2023e) align with our findings based on Twitter data. This confirms the reliability of our method and shows the potential of such an approach as a tool to rapidly produce information on crisis scenarios. In addition, we identified topics that were not present in the ground survey results. This might be due to a more open and natural communication of migrants on social media compared to predefined topics in survey questions. This reveals relevant issues that might not have been anticipated in survey designs. Nevertheless, it is important to consider that the assessment of the migrants' needs is challenging, and each data source presents its own benefits and its own limitations. Social media assessments are limited to the social network users (e.g., Twitter users, Facebook users). In the case of surveys, too, it is not possible to distribute them among all migrants, and achieving representativeness in the sample is not easy either, nor is every person keen to respond to questions related to their personal situation (and they may skip the question or answer in a way they think the researcher expects them to). In many cases, surveys are limited to specific places (normally border crossings, official reception facilities and urban centers) and based on a random sample, so they are not representative of the whole population of migrants (UNHCR, 2023e). Nevertheless, both data sources present crucial information to respond to crises and supplement each other well. Their integration provides a powerful approach to understand better patterns of migration and the needs of migrants at an early stage of a humanitarian crisis. Surveys provide a more complete picture of a migrant's personal background (e.g., ethnicity, age, spoken languages, gender, family members) and structured information related to their needs, while social media data show first-hand reports, opinions, emotions and needs in real-time from the personal perspective of the people affected by the crisis (Zhu et al., 2022).

The assessment of the content that users share on social media offers an opportunity for several applications to respond successfully to a crisis. In this research we showed the benefits of using Twitter data to identify the mobility of migrants and the topics they discussed at different stages during migration. The thematically structured information derived from this approach can be further analyzed. There is potential to quantify and map damaged infrastructure in Ukraine based on information included in topics such as "attack reports" (Lanclos & Cottray, 2022). Our results also allow to monitor the main needs of migrants over space and time. Here, it is possible to empirically differentiate between personal and common needs related to different issues, such as finances, health, politics, communication, among others. With our approach, the attitudes of the migrants expressed towards the topics identified can be further subclassified linguistically into emotions, judgements and statements of appreciation (see Martin & White, 2005 for a suitable framework). As a next step, we will implement a fine-grained linguistic analysis of the Tweets that have so far been classified into different topics at diverse migration stages, to assess further the personal experiences of the war scenario. This information is valuable to also understand better the migrants' experience from a psychological point of view, which may be used to provide them with proper targeted health assistance in future crises.

4.3. Challenges of social media datasets and how they were addressed in this study

The most evident limitation of social media data is the sampling population, which is restricted to the (active) users of the corresponding social network or social media platform (e.g., Twitter users; Zhu et al., 2022). Nevertheless, in other research as well as ours, it has been found that assessments based on social media data provide plausible and robust insights on diverse social dynamics such as human mobility and complement official information well (see for instance Hübl et al. (2017), Leasure et al. (2022a), Mazzoli et al (2020).

The presence of bots also represents a challenge when the assessment of content produced by humans is intended (Rodríguez-Ruiz et al., 2020; Schuchard et al., 2019). Moreover, it can be challenging to interpret the content and identify the function of short text messages produced in natural language containing elements such as emojis, hashtags, and URL's (Pota et al., 2021). The computation of the statistics on the Twitter activity of the users in our database showed accounts with uncommon behavior in terms of the follower-to-following ratio and the total and the average number of posts per day. Consequently, we removed potentially automated accounts representing ~24 % of the total number of Twitter users who were represented in our database. The text filter efficiently removed Tweets with content that we did not aim to analyze, such as multiple Tweets containing the same text or posts just describing the publication of pictures or videos. Moreover, removing the URL's, hashtags and emojis allowed us to define the language of "undetermined" Tweets and integrate them to our assessment.

Data privacy is another key aspect when working with social media data. Implementing the "humans as sensors" approach means that the geolocation of social media users is tracked and the discourses they share (e.g., opinions, needs, concerns) are assessed (Kounadi & Resch, 2018). Our approach is based on ethical principles; that is to say that our publications do not include coordinates or dates revealing specifically when and where individual users were located. In line with Twitter policies, and following Geoprivacy recommendations (Kounadi & Resch, 2018), none of our publications will reveal the metadata information of Tweets or Twitter users such as Twitter id, user id, user name, or user description.

In recent years Twitter became one of the main data sources for academic research based on social media data, due to its global scope, the valuable information that it includes (e.g., on geolocation), and its free access (Häberle, Hoffmann, & Zhu, 2022; Zhu et al., 2022). Researchers were able to make numerous global, regional, and local assessments thanks to the data accessibility provided by Twitter through its API. Additionally, Twitter provided an academic research access, which allowed researchers to access past and current Twitter data for non-profit purposes (Ledford, 2023). At the beginning of 2023, Twitter announced the end of the free access to its data. This change may impact ongoing and future research. For one thing, while the data is still available to those who can pay, the added value of Twitter data to e.g., produce real-time information related to wars, floods, earthquakes, and other natural and social disasters may not be exploited in many cases, as a quick download and sifting through new data may no longer be feasible. This scenario will in particular affect research groups from countries with low research budgets, which do not have the resources to pay for the Twitter API access, increasing the already high levels of scientific inequality between developed and developing economies (Ledford, 2023).

Social media represents a rich data source for research as well as for decision-makers e.g., in crisis response situations, providing first-hand information in real-time of various facets of human behavior, including needs, opinions, interests, sentiments, and in some cases, mobility (Hübl et al., 2017; Mast et al., 2023; Zhu et al., 2022). This research demonstrates that the challenges associated with social media data can be effectively mitigated through multiple methods, thereby reducing potential biases. Furthermore, the comparison of results derived from social media with other datasets (e.g., official statistics) allows to confirm the plausibility of the insights derived from such data. Although the approach presented in this research focused on Twitter, the implemented methods can be transferred to social media data from other platforms, overcoming the current lack of free access to Twitter data.

5. Conclusions

This research has joined the collective effort among researchers to produce relevant information on the ongoing humanitarian crisis triggered by the current war in Ukraine. Our aim was to identify migration patterns and the main interests, opinions and needs of migrants from Ukraine at diverse migration stages based on social media data. We integrated spatial analysis and NLP techniques to assess multilingual geolocated data produced by Twitter users who left Ukraine after the beginning of the war.

Our results reveal the suitability of such an approach to identify the main countries and cities to which migrants moved over time, as well as their main interests expressed in multiple languages at different migration stages. The plausibility of our results is confirmed by the consistency of the spatial migration patterns and migrants' needs with those stated in official sources, despite the reduced sample population in our data. In particular, our results have the potential to complement official statistics with the migrants' personal perspective recorded at various points in time, at diverse locations and at different stages of the migration process.

This approach allowed us to include the voices of migrants in real-time based on first-hand information, showing where and when they have specific interests and needs, and face particular risks. As a next step, we plan to conduct a fine-grained linguistic analysis of the Tweets that have been classified into different topics at diverse migration stages. With that, we aim to further assess the personal experiences of migrants in the ongoing war. This type of information may aid rapid response plans in future crises, taking advantage of the current massive flow of information of diverse types (e.g., earth observation data, social media data, news media) and fast-evolving fields such as NLP and geolingual studies.

CRediT authorship contribution statement

Richard Lemoine-Rodríguez: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Johannes Mast:** Methodology, Software, Writing – review & editing. **Martin Mühlbauer:** Data curation, Writing – review & editing. **Nico Mandery:** Software. **Carolin Biewer:** Conceptualization, Funding acquisition, Methodology, Resources, Writing – review & editing. **Hannes Taubenböck:** Conceptualization, Funding acquisition, Methodology, Resources, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

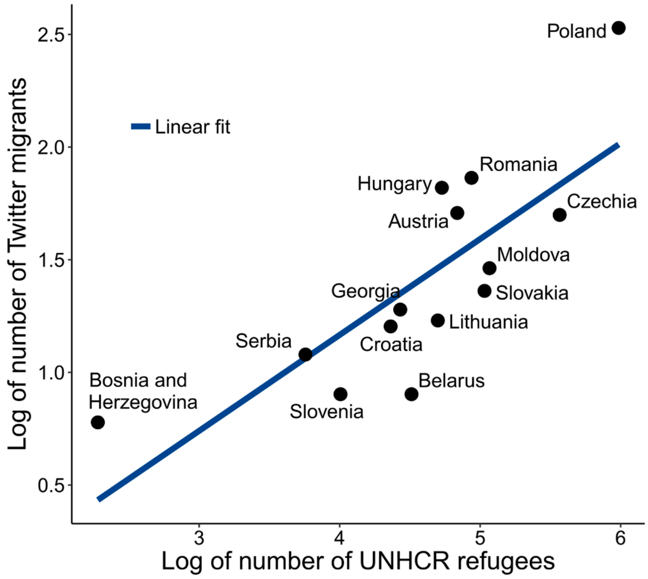
Funding

The study was conducted as part of the project "A New Focus in English Linguistics: Geolingual Studies", funded by the Volkswagen Foundation (grant number: 98 662). The study was also supported by the "Satellite based GEOinformation in urban CONflict regions for supporting humanitarian activities (GEOCON)" project of the German Aerospace Center (DLR).

Acknowledgements

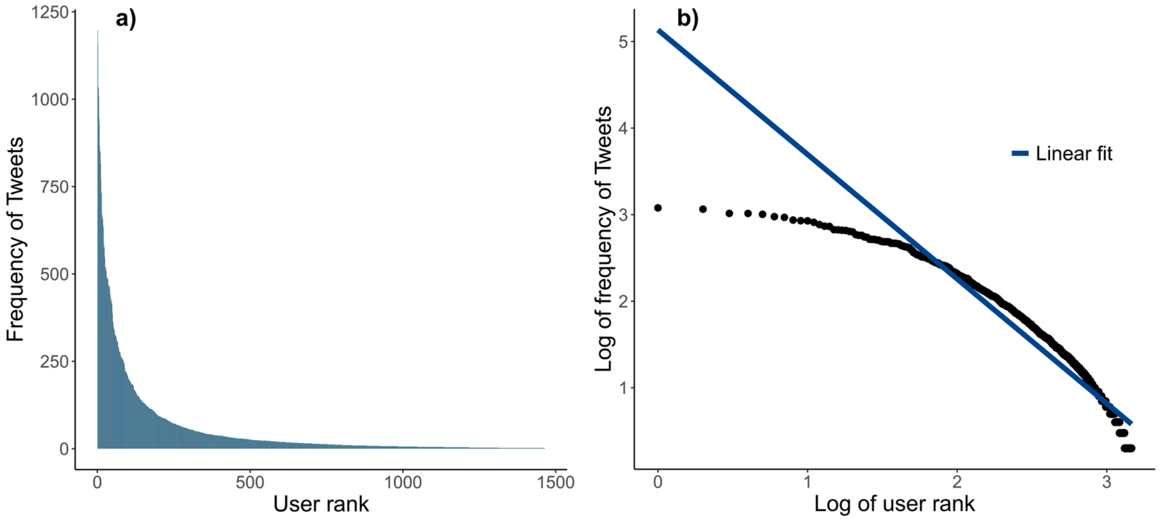
We are grateful for the academic and streaming data access provided by Twitter, which allowed us to conduct this research. The authors would also like to thank Moritz Rösch and Sebastian Düker for their assistance in data processing.

Appendix A



A comparison between the number of migrants detected via Twitter and the numbers listed in the official records of refugees by country from the UNHCR in September 2022. Only countries which were completely covered by our bounding box and in which there were Twitter users present in September 2022 are included. Values in the scatterplot are represented in log-log scale due to the different orders of magnitude of the data. A linear fit between the original values from the two sources produces an $R^2 = 0.89$.

Appendix B



Bar plot of the frequency of Tweets and users' rank based on the number of Tweets they posted (a) and its corresponding log-log plot (b).

Appendix C

Table 3
The number of Tweets per topic and migration stage.

Topic	Migration stage	Frequency
Attack reports	Before leaving	777
Attack reports	After leaving	128
Attack reports	After returning	361
Food	Before leaving	309
Food	After leaving	239
Food	After returning	253
Social media	Before leaving	323
Social media	After leaving	181
Social media	After returning	208
Art	Before leaving	147
Art	After leaving	124
Art	After returning	194
Finance	Before leaving	201
Finance	After leaving	102
Finance	After returning	143
Pets	Before leaving	182
Pets	After leaving	78
Pets	After returning	109
Music	Before leaving	162
Music	After leaving	86
Music	After returning	75
Int. support/NATO	Before leaving	171
Int. support/NATO	After leaving	91
Int. support/NATO	After returning	53
Transport	Before leaving	118
Transport	After leaving	92
Transport	After returning	91
Movies/videos	Before leaving	136
Movies/videos	After leaving	78
Movies/videos	After returning	75
Politics	Before leaving	131
Politics	After leaving	45
Politics	After returning	90
Literature	Before leaving	106
Literature	After leaving	72
Literature	After returning	68
Religion	Before leaving	110
Religion	After leaving	70
Religion	After returning	45
Photography	Before leaving	92
Photography	After leaving	53
Photography	After returning	73
Donations to Ukrainians	Before leaving	97
Donations to Ukrainians	After leaving	30
Donations to Ukrainians	After returning	88
Languages	Before leaving	106
Languages	After leaving	44
Languages	After returning	40
Health care	Before leaving	90
Health care	After leaving	40
Health care	After returning	60
Sports	Before leaving	65
Sports	After leaving	32
Sports	After returning	70
Russian government	Before leaving	84
Russian government	After leaving	36
Russian government	After returning	46
Clothes	Before leaving	67
Clothes	After leaving	38
Clothes	After returning	40
Love	Before leaving	44
Love	After leaving	70
Love	After returning	31

(*continued on next page*)

Table 3 (*continued*)

Topic	Migration stage	Frequency
Vegetation	Before leaving	61
Vegetation	After leaving	35
Vegetation	After returning	48
Humanitarian aid	Before leaving	81
Humanitarian aid	After leaving	16
Humanitarian aid	After returning	45
Food export/production	Before leaving	60
Food export/production	After leaving	20
Food export/production	After returning	56
Nazism/fascism	Before leaving	76
Nazism/fascism	After leaving	33
Nazism/fascism	After returning	21
Ukrainian places	Before leaving	74
Ukrainian places	After leaving	14
Ukrainian places	After returning	41
Nuclear threat	Before leaving	61
Nuclear threat	After leaving	24
Nuclear threat	After returning	40
Foreigners leaving Ukraine	Before leaving	58
Foreigners leaving Ukraine	After leaving	53
Foreigners leaving Ukraine	After returning	3
Job search	Before leaving	41
Job search	After leaving	23
Job search	After returning	45
Encouraging Ukraine	Before leaving	60
Encouraging Ukraine	After leaving	21
Encouraging Ukraine	After returning	26
War journalism	Before leaving	59
War journalism	After leaving	12
War journalism	After returning	33

Appendix D

Table 4
The number of Tweets per topic and language.

Topic	Language	Frequency
Attack reports	English	981
Attack reports	Ukrainian	164
Attack reports	Russian	95
Attack reports	Polish	26
Food	English	164
Food	Ukrainian	510
Food	Russian	116
Food	Polish	11
Social media	English	231
Social media	Ukrainian	400
Social media	Russian	78
Social media	Polish	3
Art	English	418
Art	Ukrainian	30
Art	Russian	14
Art	Polish	3
Finance	English	139
Finance	Ukrainian	254
Finance	Russian	49
Finance	Polish	4
Pets	English	80
Pets	Ukrainian	235
Pets	Russian	50
Pets	Polish	4
Music	English	109
Music	Ukrainian	173
Music	Russian	32
Music	Polish	9
Int. support/NATO	English	239

(*continued on next page*)

Table 4 (*continued*)

Topic	Language	Frequency
Int. support/NATO	Ukrainian	62
Int. support/NATO	Russian	14
Transport	English	120
Transport	Ukrainian	124
Transport	Russian	31
Transport	Polish	26
Movies/videos	English	84
Movies/videos	Ukrainian	159
Movies/videos	Russian	40
Movies/videos	Polish	6
Politics	English	125
Politics	Ukrainian	119
Politics	Russian	17
Politics	Polish	5
Literature	English	50
Literature	Ukrainian	155
Literature	Russian	35
Literature	Polish	6
Religion	English	58
Religion	Ukrainian	125
Religion	Russian	34
Religion	Polish	8
Photography	English	77
Photography	Ukrainian	105
Photography	Russian	33
Photography	Polish	3
Donations to Ukrainians	English	193
Donations to Ukrainians	Ukrainian	16
Donations to Ukrainians	Russian	4
Donations to Ukrainians	Polish	2
Languages	English	58
Languages	Ukrainian	111
Languages	Russian	17
Languages	Polish	4
Health care	English	25
Health care	Ukrainian	138
Health care	Russian	20
Health care	Polish	7
Sports	English	47
Sports	Ukrainian	107
Sports	Russian	11
Sports	Polish	2
Russian government	English	109
Russian government	Ukrainian	23
Russian government	Russian	34
Clothes	English	30
Clothes	Ukrainian	88
Clothes	Russian	24
Clothes	Polish	3
Love	English	55
Love	Ukrainian	57
Love	Russian	33
Vegetation	English	33
Vegetation	Ukrainian	91
Vegetation	Russian	18
Vegetation	Polish	2
Humanitarian aid	English	120
Humanitarian aid	Ukrainian	10
Humanitarian aid	Russian	2
Humanitarian aid	Polish	10
Food export/production	English	106
Food export/production	Ukrainian	23
Food export/production	Russian	4
Food export/production	Polish	3
Nazism/fascism	English	61
Nazism/fascism	Ukrainian	25
Nazism/fascism	Russian	43
Nazism/fascism	Polish	1
Ukrainian places	English	81
Ukrainian places	Ukrainian	29

(*continued on next page*)

Table 4 (*continued*)

Topic	Language	Frequency
Ukrainian places	Russian	17
Ukrainian places	Polish	2
Nuclear threat	English	102
Nuclear threat	Ukrainian	11
Nuclear threat	Russian	10
Nuclear threat	Polish	2
Foreigners leaving Ukraine	English	103
Foreigners leaving Ukraine	Ukrainian	7
Foreigners leaving Ukraine	Russian	4
Job search	English	38
Job search	Ukrainian	58
Job search	Russian	11
Job search	Polish	2
Encouraging Ukraine	English	57
Encouraging Ukraine	Ukrainian	41
Encouraging Ukraine	Russian	8
Encouraging Ukraine	Polish	1
War journalism	English	90
War journalism	Ukrainian	11
War journalism	Russian	1
War journalism	Polish	2

References

Arcila-Calderón, C., Sánchez-Holgado, P., Quintana-Moreno, C., Amores, J.-J., & Blanco-Herrero, D. (2022). Hate speech and social acceptance of migrants in Europe: Analysis of tweets with geolocation. *Comunicar, 30*, 21–35. https://doi.org/10.3916/C71-2022-02

Best, L., & Menkhoff, M. (2022). The EU's social connectedness to Ukraine and its implications for the distribution of Ukrainian refugees. *CESifo Forum, 23*, 28–35.

Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C., & West, G. (2007). Growth, innovation, scaling, and the pace of life in cities.

Bouma, G. (2009). Normalized (Pointwise) mutual information in collocation extraction. *Proceeding GSCL, 30*, 31–40.

Chi, Guanghua, Lin, F., Chi, Guangqing, & Blumenstock, J. (2020). A general approach to detecting migration events in digital trace data. *PloS One, 15*(10), Article e0239408. https://doi.org/10.1371/journal.pone.0239408

CNN. (2022). *Russia launches military attack on Ukraine with reports of explosions and troops crossing border*. https://edition.cnn.com/2022/02/23/europe/russia-ukraine-putin-military-operation-donbas-intl-hnk/index.html.

Conflict Observatory, (2023). A central hub to capture, analyze, and make widely available evidence of Russia-perpetrated war crimes and other atrocities in Ukraine [WWW Document]. https://hub.conflictobservatory.org/portal/apps/sites/#/home/.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. https://doi.org/10.48550/arXiv.1810.04805.

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology, 7*. https://doi.org/10.3389/fsoc.2022.886498

Elmqvist, T., Andersson, E., McPhearson, T., Bai, X., Bettencourt, L., Brondizio, E., et al. (2021). Urbanization in and for the Anthropocene. *NPJ Urban Sustainability, 1*, Article 6. https://doi.org/10.1038/s42949-021-00018-w

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

Häberle, M., Hoffmann, E. J., & Zhu, X. X. (2022). Can linguistic features extracted from geo-referenced tweets help building function classification in remote sensing? *ISPRS Journal of Photogrammetry and Remote Sensing, 188*, 255–268. https://doi.org/10.1016/j.isprsjprs.2022.04.006

Hübl, F., Cvetojevic, S., Hochmair, H., & Paulus, G. (2017). Analyzing refugee migration patterns using geo-tagged tweets. *ISPRS International Journal of Geo-Information, 6*(10), 302. https://doi.org/10.3390/ijgi6100302

IOM. (2022). *Ukraine returns report September 2022*.

Khatua, A., & Nejdl, W. (2021). Struggle to Settle down! Examining the Voices of Migrants and Refugees on Twitter Platform. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work CSCW* (pp. 95–98). Association for Computing Machinery. https://doi.org/10.1145/3462204.3481773.

Kounadi, O., & Resch, B. (2018). A geoprivacy by design guideline for research campaigns that use participatory sensing data. *Journal of Empirical Research on Human Research Ethics: JERHRE, 13*, 203–222. https://doi.org/10.1177/1556264618759877

Lanclos, R., & Cottray, O. (2022). Ukraine: The HALO trust maps landmines and explosive remnants of war [WWW Document]. https://www.esri.com/about/newsroom/blog/halo-trust-maps-ukraine-explosive-remnants/.

Lau, J. H., Newman Google, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 530–539).

Leasure, D. R., Kashyap, R., Rampazzo, F., Dooley, C. A., Elbers, B., Bondarenko, M., et al. (2022a). Nowcasting daily population displacement in Ukraine through social media advertising data. *SocArXiv, 49*(2), 231–254. https://doi.org/10.31235/osf.io/6j9wq

Leasure, D. R., Kashyap, R., Rampazzo, F., Elbers, B., Dooley, C., & Weber, I. et al., (2022b).Ukraine Crisis: Monitoring population displacement through social media activity.

Ledford, H. (2023). Researchers scramble as Twitter plans to end free data access. *Nature, 614*, 602–603. https://doi.org/10.1038/d41586-023-00460-z

Li, Z., Huang, X., Hu, T., Ning, H., Ye, X., Huang, B., et al. (2021). ODT FLOW: Extracting, analyzing, and sharing multi-source multi-scale human mobility. *PloS one, 16*(8), Article e0255259. https://doi.org/10.1371/journal.pone.0255259

Martin, J., & White, P. R. (2005). *The language of evaluation*. New York: Palgrave Macmillan.

Mast, J., Sapena, M., Mühlbauer, M., Biewer, C., & Taubenböck, H. (2023). The migrant perspective: Measuring migrants' movements and interests using geolocated tweets. *Population Space and Place.* , Article e2732. https://doi.org/10.1002/psp.2732

Mazzoli, M., Diechtiareff, B., Tugores, A., Wives, W., Adler, N., Colet, P., et al. (2020). Migrant mobility flows characterized with digital data. *PloS One, 15*, Article e0230264. https://doi.org/10.1371/journal.pone.0230264

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Ooms, J. (2022a). cld2: Google's Compact Language Detector 2.

Ooms, J. (2022b). cld3: Google's Compact Language Detector 3.

Orabi, M., Mouheb, D., Al Aghbari, Z., & Kamel, I. (2020). Detection of bots in social media: A systematic review. *Inf. Process. Manag., 57*, Article 102250. https://doi.org/10.1016/j.ipm.2020.102250

Pota, M., Ventura, M., Catelli, R., & Esposito, M. (2021). An effective bert-based pipeline for twitter sentiment analysis: A case study in Italian. *Sensors (Switzerland), 21*, 1–21. https://doi.org/10.3390/s21010133

R Core Team. (2020). *R: A language and environment for statistical computing*.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

Rodríguez-Ruiz, J., Mata-Sánchez, J. I., Monroy, R., Loyola-González, O., & López-Cuevas, A. (2020). A one-class classification approach for bot detection on Twitter. *Computers & Security, 91*, Article 101715. https://doi.org/10.1016/j.cose.2020.101715

Roesslein, J. (2020). *Tweepy: Twitter for Python!*. URL: https://Github.Com/Tweepy/Tweepy.

Schuchard, R., Crooks, A. T., Stefanidis, A., & Croitoru, A. (2019). Bot stamina: Examining the influence and staying power of bots in online social networks. *Applied Network Science, 4*, Article 55. https://doi.org/10.1007/s41109-019-0164-x

Senaratne, H., Mühlbauer, M., Götzer, S., Riedlinger, T., & Taubenböck, H. (2023). Detecting crisis events from unstructured text data using signal words as crisis determinants. *Internatiional Journal of Digital Earth, 16*(2), 4601–4620.

Shamoug, A. (2022). In Ukraine, machine-learning algorithms and big data scans used to identify war-damaged infrastructure [WWW Document]. https://www.undp.org/blog/ukraine-machine-learning-algorithms-and-big-data-scans-used-identify-war-damaged-infrastructure.

Smith, L. G. E., McGarty, C., & Thomas, E. F. (2018). After Aylan Kurdi: How tweeting about death, threat, and harm predict increased expressions of solidarity with refugees over time. *Psychological Science, 29*, 623–634. https://doi.org/10.1177/0956797617741107

Twitter. (2022). *How to detect signal from noise and build powerful filtering rules [WWW Document]*. URL https://developer.twitter.com/en/docs/tutorials/building-powerful-enterprise-filters/example accessed 6.11.22.

UNHCR. (2022a). *Profiles, needs & intentions of refugees from Ukraine*.

UNHCR. (2022b). *Ukraine situation flash update No 31*.

UNHCR. (2022c). *Ukraine situation flash update No 6*.

UNHCR. (2023a). *Ukraine situation flash update No 40*.

UNHCR. (2023b). *Ukraine situation regional refugee response plan*.

UNHCR. (2023c). *Ukraine situation flash update No 58*.

UNHCR. (2023d). *Protection risks and needs of refugees from Ukraine*.

UNHCR. (2023e). *Profiles, needs & intentions of refugees from Ukraine*. https://data.unhcr.org/en/situations/ukraine.

Vahdat-Nejad, H., Ghasem Akbari, M., Salmani, F., Azizi, F., & Nili-Sani, H.-.R. (2023). Russia-Ukraine war: Modeling and clustering the sentiments trends of various countries.

Zhu, X. X., Wang, Y., Kochupillai, M., Werner, M., Haberle, M., Hoffmann, E. J., et al. (2022). Geoinformation Harvesting From Social Media Data: A community remote sensing approach. *IEEE Geoscience and Remote Sensing Magazine, 10*, 150–180. https://doi.org/10.1109/MGRS.2022.3219584