International Journal of Geo-Information

MDPI

Article

Spatio-Temporal Machine Learning Analysis of Social Media Data and Refugee Movement Statistics

Clemens Havas [1,*], Lorenz Wendlinger [2], Julian Stier [2], Sahib Julka [2], Veronika Krieger [1], Cornelia Ferner [3], Andreas Petutschnig [1], Michael Granitzer [2], Stefan Wegenkittl [3] and Bernd Resch [1,4]

1 Department of Geoinformatics, University of Salzburg, 5020 Salzburg, Austria; veronika.krieger@sbg.ac.at (V.K.); andreas.petutschnig@sbg.ac.at (A.P.); bernd.resch@sbg.ac.at (B.R.)
2 Department of Data Science, University of Passau, 94032 Passau, Germany; lorenz.wendlinger@uni-passau.de (L.W.); julian.stier@uni-passau.de (J.S.); sahib.julka@uni-passau.de (S.J.); michael.granitzer@uni-passau.de (M.G.)
3 Information Technology and Systems Management, Salzburg University of Applied Sciences, 5412 Puch, Austria; cornelia.ferner@fh-salzburg.ac.at (C.F.); stefan.wegenkittl@fh-salzburg.ac.at (S.W.)
4 Center for Geographic Analysis, Harvard University, Cambridge, MA 02138, USA
* Correspondence: clemensrudolf.havas@sbg.ac.at

Abstract: In 2015, within the timespan of only a few months, more than a million people made their way from Turkey to Central Europe in the wake of the Syrian civil war. At the time, public authorities and relief organisations struggled with the admission, transfer, care, and accommodation of refugees due to the information gap about ongoing refugee movements. Therefore, we propose an approach utilising machine learning methods and publicly available data to provide more information about refugee movements. The approach combines methods to analyse the textual, temporal and spatial features of social media data and the number of arriving refugees of historical refugee movement statistics to provide relevant and up to date information about refugee movements and expected numbers. The results include spatial patterns and factual information about collective refugee movements extracted from social media data that match actual movement patterns. Furthermore, our approach enables us to forecast and simulate refugee movements to forecast an increase or decrease in the number of incoming refugees and to analyse potential future scenarios. We demonstrate that the approach proposed in this article benefits refugee management and vastly improves the status quo.

Keywords: spatio-temporal; machine learning; refugee movements; simulation; forecasting; social media

Citation: Havas, C.; Wendlinger, L.; Stier, J.; Julka, S.; Krieger, V.; Ferner, C.; Petutschnig, A.; Granitzer, M.; Wegenkittl, S.; Resch, B. Spatio-Temporal Machine Learning Analysis of Social Media Data and Refugee Movement Statistics. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 498. https://doi.org/10.3390/ijgi10080498

Academic Editor: Wolfgang Kainz

Received: 8 June 2021
Accepted: 20 July 2021
Published: 23 July 2021

1. Introduction

Ongoing conflicts in the Near East led to mass refugee movements to Central Europe in 2015 and 2016 when more than 2.5 million people applied for asylum in member states of the European Union [1]. One of the primary paths to Europe was the informal "Balkan route" that spans from Turkey to Central Europe through Greece, North Macedonia, Albania, Kosovo, Montenegro, Serbia, Bosnia and Herzegovina, Slovenia, and Hungary. From 2015 until March 2016, 1.2 million people moved along this route. In this period, the refugee movements led to challenging situations in the affected countries as the movements along the Balkan route did not result from a coordinated decision among these countries but instead developed gradually [2]. Local public authorities and relief organisations in these countries faced the extreme logistical challenge of providing for such a large number of arriving refugees. Besides these practical problems, public authorities and relief organisations had to deal with a lack of information about refugee movements, making the organisation of efficient and effective humanitarian aid virtually impossible [3]. Especially at the beginning of 2015, refugee groups of varying sizes arrived at the border at all hours of the day without prior notification. This disorganisation was caused by the limited communication between those involved and the inability to collect information. It

demonstrated the need for accurate information that can be collected quickly and at any time throughout humanitarian crisis situations.

In the big data era, addressing and solving the aforementioned problems is a necessary task that should be a high priority for all involved authorities and organisations. For one, there are various data sources with different characteristics available at any time that include useful information about the current state of refugee movements. Human sensors are of particular interest and comprise social media networks (sharing publicly useful information), crowdsourcing (completing online tasks) or offline solutions for pedestrians (completing offline tasks) [4]. Social media have proven to be useful as they provide timely big data across borders and can be used for real-time monitoring of collective refugee movements. Furthermore, refugees commonly use social media networks as a platform to communicate and collect information [5]. Thereby, for the data collection and analysis the principles of "Collective Sensing" can be followed [6].

Other human sensor datasets could serve as further major data sources such as high temporal and spatial resolution data of refugees moving between places within countries, but could not be obtained for this study. Therefore, we substitute human sensor datasets with refugee movement statistics supplied by the United Nations High Commissioner for Refugees (UNHCR), which provides statistics about the number of arriving refugees at a temporal resolution of one month. This UNHCR dataset can be used to analyse past refugee movements and serve as the basis for statistical models to forecast and simulate refugee movements. Social media data and the UNHCR dataset complement each other as social media enable real-time monitoring and the UNHCR data allow us to learn from past trends and thus predict future trends.

In this publication, we propose an approach on how social media and UNHCR datasets can be analysed with state-of-the-art machine learning methods based on multiple features. This includes extracting information about current refugee movements, forecasting daily arrivals per country, and simulating refugee movements in a network model that goes beyond current research efforts. The purpose is to provide relief organisations and public authorities with a sophisticated approach to have access to more information during large-scale refugee movements.

The paper starts with reviewing related work. Then, we lay the foundation of our analysis by providing exploratory analysis results on refugee-related Twitter data (tweets) to show that the collected Twitter dataset can be used to capture real-world events. For this research, we acquired tweets through spatial and temporal queries before filtering the dataset with keywords and performing a spatio-temporal analysis on this subset. Through this, we aimed to detect collective refugee movements and hot spots. Next, we trained a Convolutional Neural Network (CNN) to identify tweets describing refugee movements, including various facts mentioned in the tweet text (refugee-related keywords, origin, destination, means of transport and number of refugees). Additionally, we applied statistical models on the UNHCR dataset to forecast the number of incoming refugees per country. We assume that the UNHCR data could easily be collected by Non-Governmental Organisations (NGOs) through digital media (e.g., smartphones, websites). However, especially in a situation with limited data, forecasting only provides limited insights. Consequently, we also discuss simulation approaches relying on network models in order to analyse potential future scenarios. The analysis results show the potential benefits and limitations of collecting useful information for public authorities and relief organisations, which we discuss in detail at the end of the paper. Below, we list the research questions that we aimed to answer through this study:

- RQ1: How can we extract spatio-temporal patterns from refugee-related tweets that match with collective real-world refugee movements?
- RQ2: How can we identify tweets that include movement information about collective refugee movements?
- RQ3: Which time-series models are suitable for forecasting the number of incoming refugees per country?

- RQ4: How can we design a network model to simulate potential refugee movements related to real-world events?

2. Related Work

2.1. Extracting Refugee Movement Information from Social Media

Numerous articles were published related to the refugee movements to Europe in 2015/2016 that focussed on the social–economic aspects and policy-making [7–10]. For our research, articles about the common practices of refugees on their way to Europe are of particular interest.

In interviews with refugees that made their way along the Balkan route, refugees stated that one of the essential items on their journey were smartphones with which they could communicate with others [5]. They share experiences on social media and interact with people who can help them on their way, such as smugglers. Furthermore, they can retrieve and share documents about their journey and locate crucial stops along their way. Communication often takes place on social media. The most popular social media networks for refugees are Facebook, Twitter, Instagram and Google Plus [11]. However, refugees are also cautious about using social media as they fear surveillance by others, such as the police or coast guards. Therefore, they sometimes switch to encrypted channels and closed groups [12]. Other obstacles in using social media are practical issues such as problems with their smartphones or internet access.

As refugees actively use social media en route to their destination, it is a highly valuable and informative resource for refugee-related analysis and information extraction. Curry et al. [13] discuss the potential of social media data for extracting information about refugee movements by showing examples of analysed Flickr and Twitter data. Social media data that include geographical and social information about refugee movements are particularly useful and are often even used by authorities and relief organisations. However, the use of social media data is still an open research field as multiple issues such as potential biases, data quality, and data processing are yet to be solved. Hübl et al. [14] aimed to determine refugee routes based on geo-tagged tweets in the context of the refugee movement in 2015. They aggregated refugee trajectories to identify refugee migration patterns. However, due to data scarcity, they could only extract a limited number of trajectories that could be associated with refugees. Furthermore, they detected topical clusters along the routes that identified geographic areas of high refugee-related Tweet activities. Although they could extract insights about refugee patterns, they concluded that the tweets' text content should be analysed to extract more location information and enhance the analysis results. Petutschnig et al. [15] spatially analysed geo-social media data to explore the options to enhance information availability during refugee movements. They applied various methods on semantic, spatial, and temporal features of the geo-social media dataset and demonstrated that the combination of analysing multiple features shows the most promising results. Their research can be extended by analysing the text with more sophisticated methods to acquire more details about the content of the posts. They also did not use any other datasets in their analysis.

Machine learning models can identify refugee-related data based on their text by training them with annotated data. Word Embeddings are a powerful concept that allows the incorporation of context for preserving semantic similarities between words. This offers more flexibility than the traditional pipeline of stemming, lemmatisation and other pre-processing followed by n-gram generation to form bags-of-words (BoW). To capture dependencies between words, which is neglected by BoW models, algorithms that are designed for time-series data, such as Recurrent Neural Networks or Long Short Term Memory networks, can be used. Recent work by Bai et al. [16] suggests that for many sequence modelling tasks, a CNN actually outperforms those methods. It is therefore pertinent to investigate a combination of CNNs operating on word-level embeddings, as has been performed successfully for sentiment classification, question classification and topic categorisation [17–20].

2.2. Forecasting and Simulating Refugee Movements

Given the economic and social impact of large-scale refugee movements, forecasting and simulating future developments are imperative for risk minimisation. Public authorities and relief organisations can incorporate proven techniques into their system to prepare for future developments.

Simulating spatio-temporal processes can address several objectives, such as identifying connections between entities on different geo-spatial levels, detecting central points on a route, and understanding the dynamics in a geo-spatial network over time. Simulations can be performed on multiple geo-spatial levels where countries, regions or cities are represented as nodes in a network that reveals insights within a country or on an inter-country level [21–23]. By using various network measures such as centrality indices, relevant entities can be detected that can help policymakers in their decision making [22,24]. Agent-based models allow us to understand the dynamics in a network model, i.e., Lin, Carley, and Cheng [25] examine who is affected by climate change and to which countries people will migrate. Suleimenova, Bell, and Groen [26] propose a generalised simulation development approach to predict the destinations of refugee movements in conflict regions. These models can also be used to simulate refugee movements. Rossetti et al. [27] provide a technical framework for simulating diffusion processes on complex networks. Refugee movements can be seen as a diffusion process through a spatial infrastructure network. Another technical framework is created by Donges et al. [28] that includes measures and models spatially embedded networks. However, data availability issues mean it is still challenging to use these frameworks, and the models must be parametrised specifically for each use case.

A vast majority of the traditional methods for forecasting refugee movements rely on maximum likelihood estimates and can be included in the frequentist domain in forecasting [29]. Traditional time series methods, which work on univariate time series, such as autoregressive integrated moving average (ARIMA) models, fall under the category of deterministic frequentist models and are often useful when the data are rich in quality and size. These models can have a short or long memory depending on the parameters. In the past, ARIMA models have been applied to birth forecasting, international migration forecasting, and energy consumption forecasting, among others, by using estimates and their respective confidence intervals [30–32]. However, these methods can be limiting when working with high uncertainty in data and forecasted processes.

Another branch of forecasting methods known as probabilistic or stochastic models allows greater control of uncertainty and its representation [33]. A state space model integrates various constituents of a time series, such as trend, seasonality, cycle, and variation, separately and then uses them together to predict the dependent variable. The seasonal autoregressive integrated moving average (SARIMA) method, for instance, can model the seasonal component in a time series. Dynamic linear models are Bayesian-based inference models, which allow for incorporating more covariate time series. Probabilistic models assign probabilities for various outcomes to occur, based on a set of assumptions about the underlying probability distributions, such as prior knowledge on the likelihood function and expert knowledge. Other approaches include the econometric model [34], which is capable of predicting both refugee movements and verifying the underlying economic theories, and the gravity model [35], where population sizes act as masses drawing people across a spatial distance.

Bijak et al. [36] aptly point out three existing uncertainties that exist in forecasting refugee movements: (1) Uncertainty arising from data, which can be attributed to errors in the data, or missing or inconsistent entries; (2) external uncertainties, such as geopolitical and natural factors; and (3) forecasting uncertainty arising from the choice of model.

In the case of refugee movements, these uncertainties are amplified due to a lack of immediately available, accurate and complete data and the high volatility of policy changes in the affected areas. Further, the analysis of one temporal context is almost impossible to extrapolate to another temporal context. Due to these challenges, the refugee movement

forecasts are highly prone to errors. The optimal forecasting strategy, therefore, involves designing a risk management framework based on a thorough investigation of individual time series and their respective uncertainty measurements, such as confidence intervals, to minimise risk on a case by case basis.

3. Refugee Movements 2015/2016 and Data Collection

Our research focuses on analysing the refugee movements in 2015 and 2016 from the Near East to Central Europe. For our analysis, we collected two Twitter datasets and the official numbers of arriving refugees per country from UNHCR for this period, as described in this section.

3.1. Twitter Datasets

Our study includes two Twitter datasets, the so-called Geo-Tweets and the General-Tweets. The *Geo-Tweets* were retrieved through the Twitter Streaming API and the Twitter REST API by requesting explicitly georeferenced tweets. Then, we merged our crawled dataset with the Harvard CGA Geotweet Archive [37]. We subsequently spatially and temporally filtered the dataset to the bounding box [8.0° E, 28.2° N, 43.2° E, 50.0° N] in the World Geodetic System 1984 (WGS 84), and to the time interval between January 2015 and December 2016. This dataset builds the basis for the spatio-temporal analysis in Section 4.1 and is used in Section 4.2.

The *General-Tweets* dataset is based on the online archive "Archive Team: The Twitter Stream Grab", which provides a 1% random sample of all public tweets [38,39]. This additional dataset is necessary for the semantic analysis in Section 4.3. A total of 3275 of the General-Tweets from September and October 2015 were annotated regarding their relevance in terms of containing quantitative movement information of refugees/migrants into Hungary, Austria, and Germany. They were annotated by three persons and only marked as relevant if they mentioned refugees or migrants, a quantity, a movement location, and if the information was about events in Europe. Figure 1 shows an overview of the datasets, and more details about the annotation process can be found in Urchs et al. [40].

Name	Number of Tweets	Timeframe
Geo-Tweets	97,653,736	January 2015 –December 2016
General-Tweets	239,680,779	January –December 2015

Figure 1. Characteristics of Twitter datasets.

3.2. United Nations High Commissioner for Refugees Dataset

The data used for forecasting were provided through the UNHCR archives [41]. We consider the UNHCR data as a substitute for data gathered by NGOs and others during ongoing refugee movements that can provide real-time updates on the situation. Forecasting based on these data supports decision making processes. The UNHCR dataset contains daily estimated arrivals from Italy, the Greek Islands, North Macedonia, Serbia, Croatia, Hungary, Slovenia, Austria, and departures from the mainland of Greece. The data stream includes estimates on nine sub-routes between the 1st of October 2015 and the 30th of September 2016. Two distinct modes of arrivals are reflected in the distribution of arrivals into Serbia (land arrivals) and in the distribution of arrivals into Greece (sea arrivals), visualised in Figure 2. In both time-series, the number of arrivals increases in autumn and decreases in the following months. An autocorrelation plot of all the different arrival and departure time series (Figure 3) shows a consistent pattern amongst all with high autocorrelation at short lags (1–5 days). This indicates that short-term local temporal information could be the most useful and can be considered when selecting the parameter values for a time-series model. A kernel density estimate (KDE), illustrated in Figure 4, shows that high magnitude arrivals are rare and most of the values are close to zero. Due

to this characteristic, modelling the time-series is difficult as the larger arrival numbers behave like outliers.

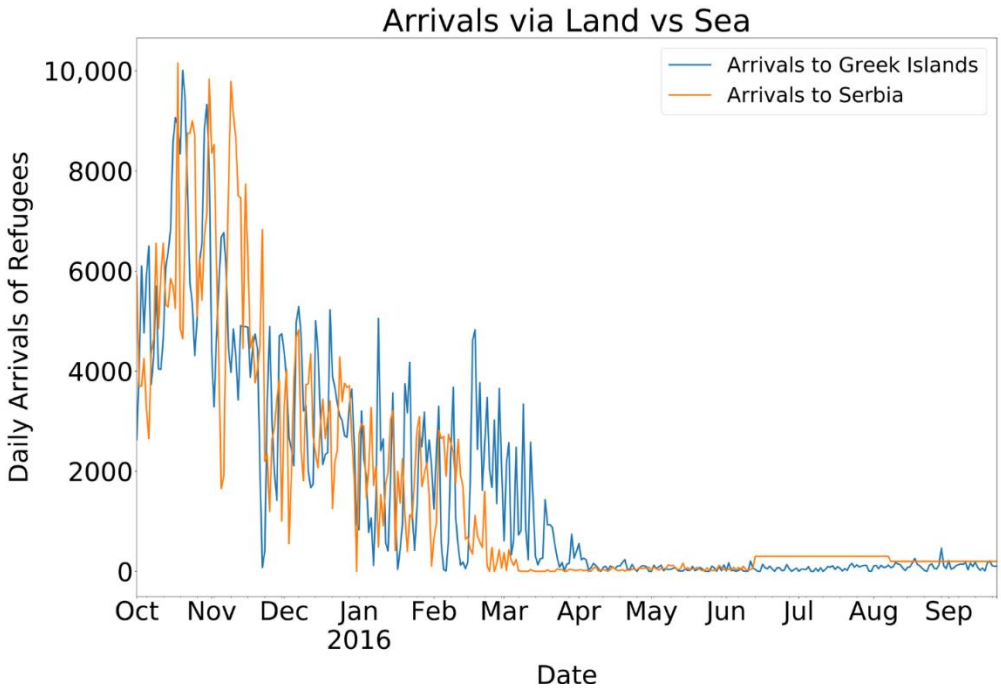

Figure 2. Overview of incoming refugee flows in Greek Islands (representing sea arrivals) and Serbia (representing land arrivals).

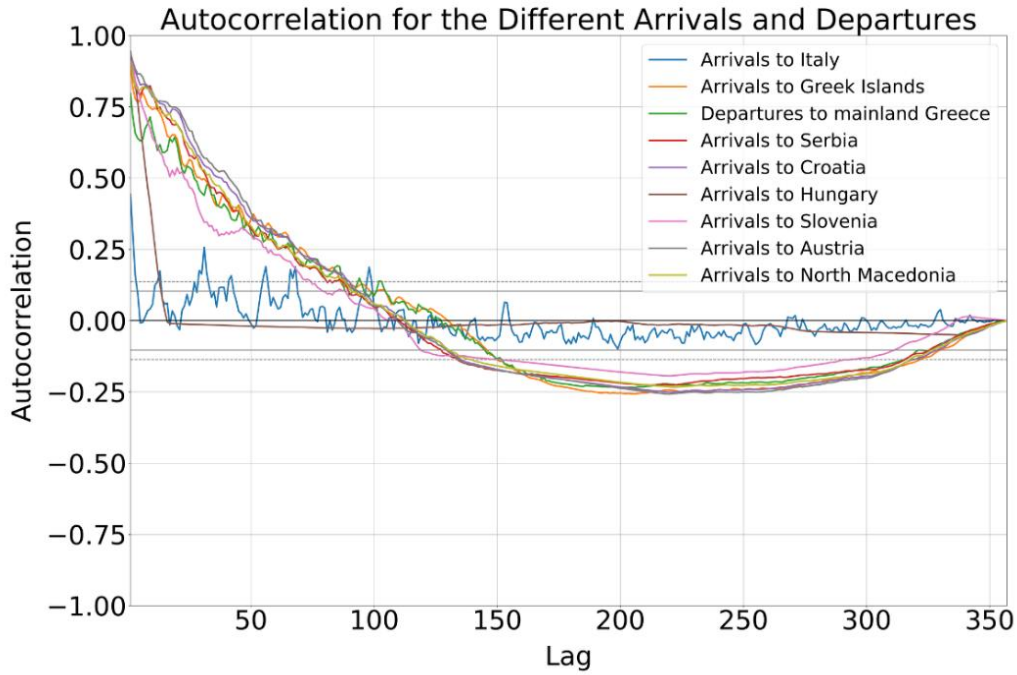

Figure 3. Autocorrelation plots for the different arrivals and departures indicating good autocorrelation functions at short lags (1–5 days).

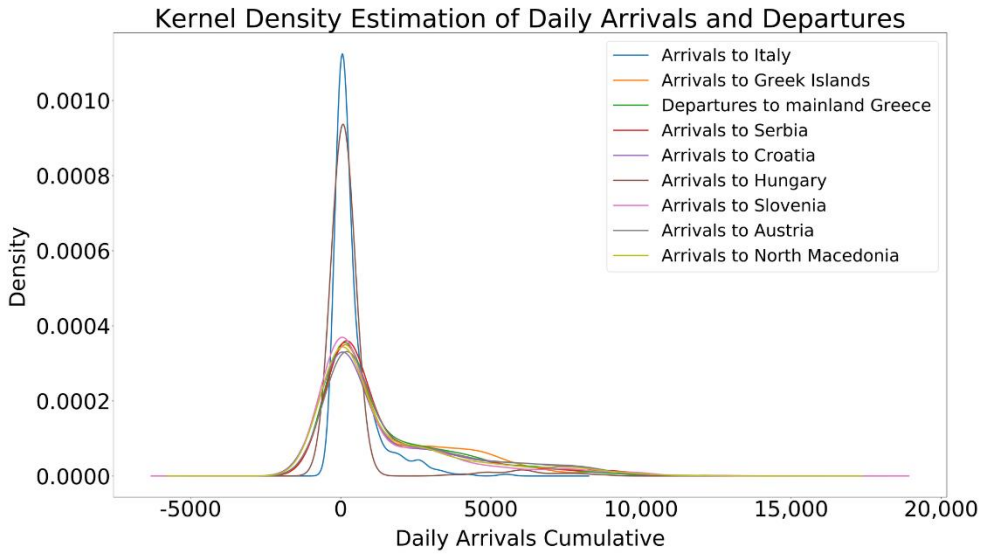

Figure 4. KDE plot showing extreme right skewing denoting the sporadicity of influx events with high magnitude.

4. Methods: Extracting, Simulating and Forecasting Refugee Movements

The approach proposed in this paper is demonstrated in Figure 5. Both Twitter datasets were filtered using expert keywords (Section 4.1), and a hot spot analysis was applied on the Geo-Tweets to extract collective refugee movement paths (Section 4.2). The keyword-filtered tweets were additionally filtered using a pre-trained CNN to extract tweets including factual information (Section 4.3). From these tweets, relevant facts for emergency organisations, such as means of transport, number of refugees, and origin–destination pairs, were identified to describe refugee movements quantitatively. Lastly, UNHCR data were used to forecast refugee movements using the ARIMA stochastic model (Section 4.4). We also designed a network model for simulation using the UNHCR dataset (Section 4.5).

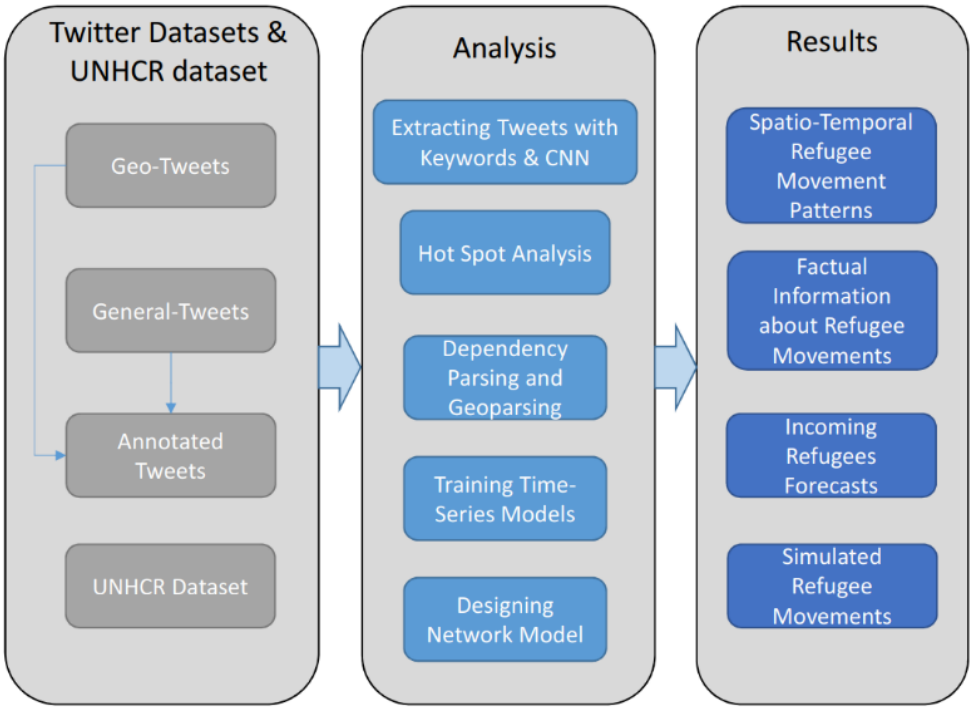

Figure 5. High-level schematic illustration of our proposed approach.

4.1. Extracting Refugee-Related Tweets with Keywords

We used keyword filtering to exclude a large number of Geo-Tweets not related to refugees. We defined a set of keywords for each of the predominantly used languages in the study area, namely English, German, Hungarian, Serbo-Croatian, Greek and Arabic. The selection of terms and translation were guided by experts from relief organisations and native speakers. The keywords include terms explicitly referring to refugees, but also terms for entities and concepts related to the refugee movement (see Figure 6).

German	English	Hungarian	Serbo Croatian	Greek	Arabic
Flüchtling	refugee	migráns	Sirija	πρόσφυγες	لاجئ
Fluechtling	migra	migrans	Сирија	προσφύγων	لاجئين
Flucht	syria	Szíria	izbeglice	πρόσφυγας	لاجئون
Migrant	border*cross	sziria	izbjeglice	μετανάστης	مهاجر
Syrien	cross*border	Soros	избеглице	μεταναστες	مهجر
Grenz		StopSoros	migranti	μετανάστες	مهجرون
Asyl		brüsszel	мигранти	Συρία	مهجرين
unbegleitete Minderjährige		bruesszel	migracije	Βρυξέλλες	لاجئة
unbegleitete Minderjaehrige		Brusszel	миграције	Prosfuga	لاجئات
Schlepper		O1G	азил	Prosfiga	مهاجرة
Willkommenskultur		kvóta	granica	Prosfyga	مهجرة
Erstaufnahmezentrum		kvota	граница	Prosfuges	مهجرات
Balkanroute		fidesz	Brisel	Prosfiges	سوريا
		bevandorlas	Брисел	Prosfyges	لاجئ
		bevándorlás	Батровци	Prosfugwn	لاجئين
		keritesepites	Batrovci	Prosfigwn	لاجئون
		kerítésépítés	Хоргош	Prosfygwn	ملجأ
		ahataron	Horgoš	Metanastis	مأوى
		ahatáron	azilanti	Metanasths	مهاجر
		migráncs	азиланти	Metanastes	مهجر
		bevandored		Metanastwn	مهجرون
		menekült		Brixelles	مهجرين
		dzsihadista		Bruxelles	حدود
		dzsihadistak		Synora	حدودية نقطة
		terrorista		Sunora	حدوديخط
		határainkat		Πρόσφυγ	معبر
		hatarainkat		Προσφύγ	لجوء
		védyük meg		Μετανάστ	أزمة
		muszlim		Μεταναστ	السورية زمةالأ
		muzulman		Prosfug	السوريةال حرب
				Prosfig	البلقانطر ق
				Prosfyg	مخيم
				Metanast	الركبان
					الزعتري

Figure 6. Refugee-related keywords in German, English, Hungarian, Serbo-Croatian, Greek and Arabic.

As a sanity check for the representativeness of these terms, we developed the following hypothesis: The refugee movement saw recurrent peaks over time, often following combat operations, which in turn lead to a rising number of tweets associated with this topic. We consider a set of keywords to be indicative of related tweets if its occurrence over time exhibits a peak at the same time as refugee numbers do. Figure 7 shows the time series of keyword matches within selected countries. We found that major events during the war appear to be reflected in the Twitter data, such as the Palmyra offensive in May 2015, which led to ISIL controlling 50% of the Syrian territory [42], or the Aleppo offensive in late 2015 that caused tens of thousands to flee the city [43].

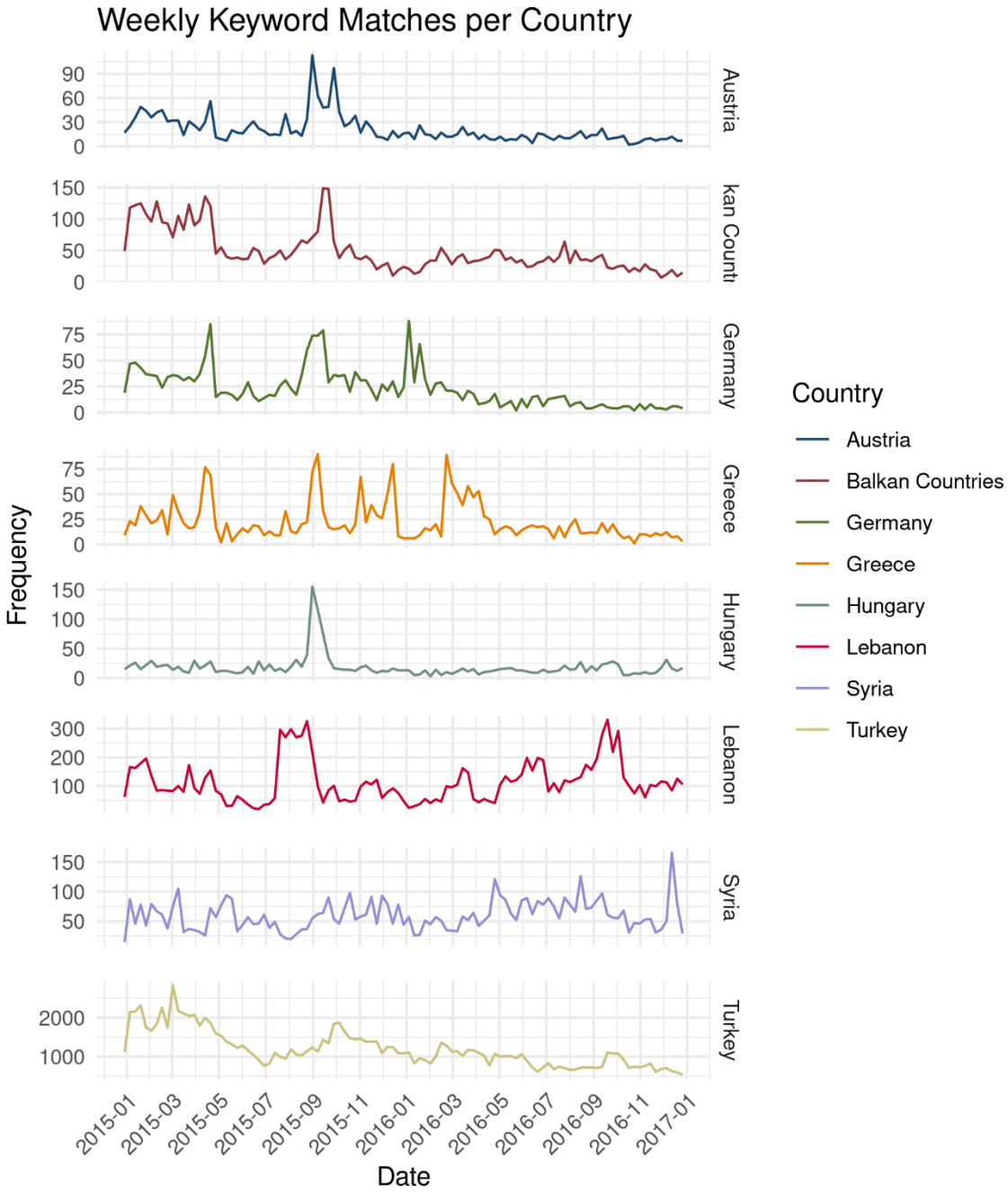

Figure 7. Development of keyword matches for the different languages over time.

To cross-check the validity, we compared the frequencies of keywords in our Twitter dataset with those on Google Trends [44]. Google Trends is a tool that visualises the frequencies of queries on the search platform. The search data are normalised by scaling the values between 0 and 100 [45]. We normalised the Twitter data to make them comparable with the Google Trends output. Figure 8 illustrates the trends for the German keyword "flüchtling" (German for "refugee") for the two media. Although a first peak in the Twitter data occurs in April 2015 and a peak in the Google Trends data occurs by the end of January 2016, both platforms exhibit a high at the beginning of September 2015, and have a similar slope in other periods.

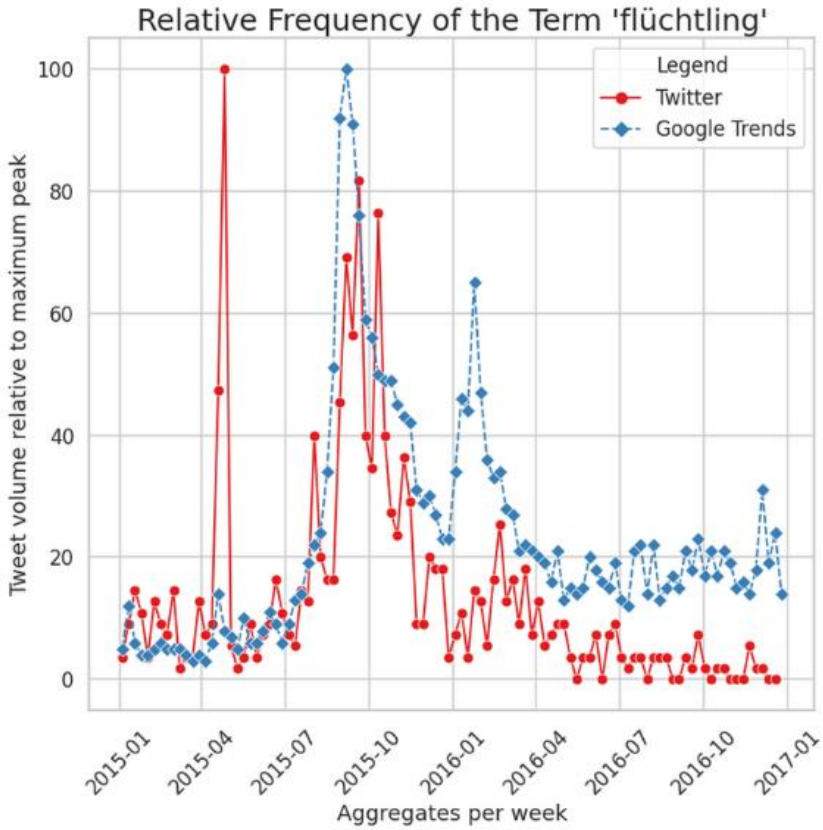

Figure 8. Relative frequency of tweets containing the German keyword "flüchtling" (red, solid line) compared to the relative frequency of search queries for that term on the Google search platform (blue, dotted line) for the Geo-Tweets.

We examined the co-occurrence of the suggested keywords in our dataset across the various languages. Figure 9 focuses on German and English keywords and exhibits higher absolute values for the term-pairs "syria" and "refugee", "syria" and "border", "border" and "refugee" and "refugee" and the stem "migra". It is worth noting that although they are English keywords, these terms also occur in German tweets, e.g., "#refugeeswelcome" [46] was a popular hashtag in Germany and Austria, and the word stem "migra" has the same meaning and application in German. The two German keywords with higher co-occurrence values are the term "flüchtling" for refugee and "flucht" signifying the act of fleeing. More specific German keywords such as "minderjährige" (referring to "unbegleitete Minderjährige", unaccompanied minor) have a smaller frequency overall.

4.2. Identifying Refugee Movements through Spatio-Temporal Clustering of Tweets

A critical task for emergency services is monitoring refugee movements across borders to prepare for incoming refugees and to estimate the time of arrival at their desired destination. Therefore, we aimed to visualise collective refugee movements on the Balkan route by applying a spatio-temporal analysis on refugee-related tweets. The analysis works under the assumption that people are more likely to comment about refugees on social media when they observe and/or are impacted by refugee movements.

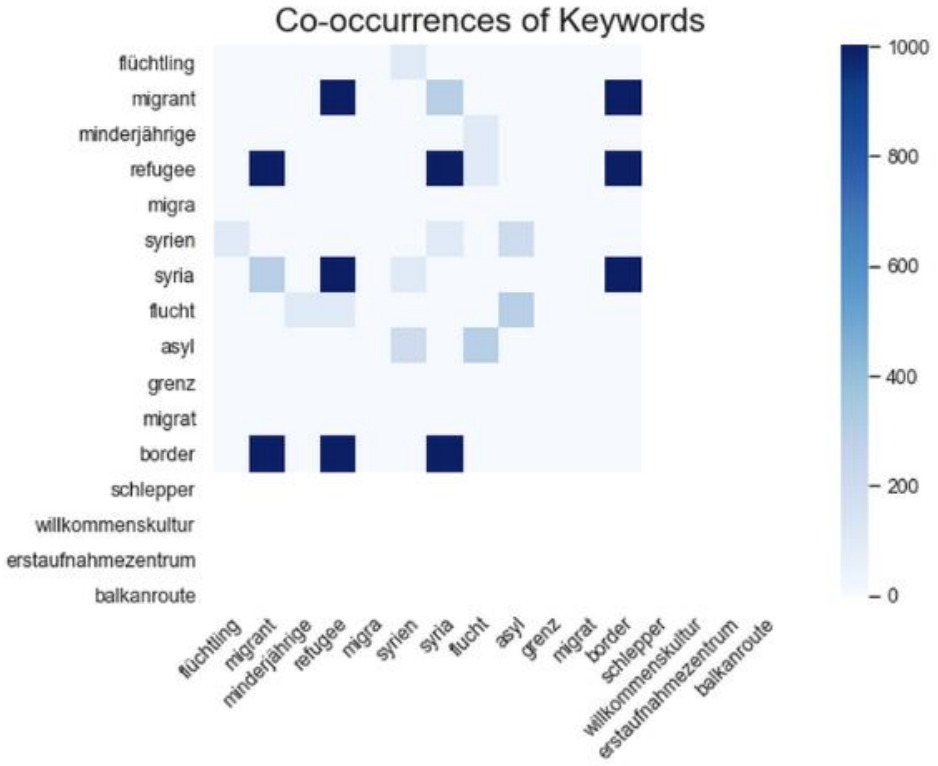

Figure 9. Absolute frequencies of German and English keyword co-occurrences in the dataset. The diagonal is set to zero to allow for better colour scaling.

For this analysis, we used the refugee-related Geo-Tweets (see Section 4.1). We binned the tweets weekly to enable the identification of collective refugee movements by comparing different points in time. We chose weekly bins as refugees oftentimes find themselves stuck at borders for many days and travel in groups in which they can cross borders at certain times. In the next step, we applied a hot spot analysis on the weekly binned refugee-related tweets to visualise spatial patterns related to refugees in the area of interest. The hot spot analysis is based on Getis-Ord G_i^*, which detects local clusters of high or low values in a spatial neighbourhood [47]. To apply a hot spot analysis on the refugee-related tweets, we created a grid of rectangular cells and aggregated the number of refugee-related and unfiltered tweets in every cell. The cell depends on the number of tweets in the area of interest and on the area's size. The calculation of the cell size is based on [48], where the cell size length is calculated as follows:

$$l = \frac{1}{d}\sqrt{2\frac{A}{n}}, \tag{1}$$

where $A\ [m^2]$ is the size of the area of interest, n is the number of refugee-related tweets and $d \in \mathbb{R}$ is used to adjust the cell size for use case specific requirements. We set d to twelve in order to have more granular cells that could be necessary to detect crossings along the border. Furthermore, we excluded cells without tweets from the analysis to differentiate between zero and null values in the cells. Subsequently, every shown cell in the hot spot maps also represents at least one tweet.

4.3. Extracting Refugee Movement Information from Tweet Text

In Section 4.2, we explain how we extracted refugee-related tweets using keywords, but further analysis is required to identify facts about refugee movements in the text. Therefore, we trained a neural network to extract tweets that include information about refugee movements following the examples in Section 3.1. The data used in this task are multi-lingual, which needs to be reflected in the design of the fact extraction methods.

We decided against using a machine translation approach, as the risk of losing valuable information through mis-translation is high in a domain with frequent mis-spellings and jargon. Some parts of the fact extraction pipeline can natively handle multiple languages, while others can be easily extended to cover additional languages. As a proof of concept, all methods were extended to work with both English and German text, as these languages are most prevalent in the labelled Twitter dataset.

The tweets should include information about the means of transport, number of refugees, origin, and destination. We used DBpedia Spotlight to recognise location names in the text [49]. Movement information could be extracted if the location names were referenced by a preposition or verb indicating movement. A distinction between the origin and destination was made by categorising the keywords in movements towards and away from a location. Finally, the coordinates for a location were determined via DBpedia [50] using the DBpedia:Locations extracted in spotting. This was performed to directly depict locations on a map or to aggregate across specific regions. We used the 2016 version of geocode Nomenclature of Territorial Units for Statistics (NUTS) to divide countries into regions. To do this, the NUTS entity that matches DBpedia:Locations in the total area was selected from all NUTS entities that contain a location.

4.3.1. Convolutional Neural Network Architecture for Identifying Refugee-Related Tweets

Our model is based on pre-trained task-specific embeddings generated in the work of Pennington et al. [51]. They consist of 200-dimensional vectors extracted from a multilingual corpus of 2 billion tweets. An advantage of this collection is the availability of specific embeddings for stop words such as the RT marker or user mentions. This collection of embeddings also encompasses tokens in all of the most frequent languages of tweets, thereby implicitly providing multilinguality support.

In our analyses, we tried various neural network settings to achieve the highest average precision (AP) [52] value in 10-fold cross-validation in the training set. The architecture is comprised of three parallel convolutional blocks with filter sizes of (1, 200), (3, 200), and (5, 200) and 192 filters each. The features extracted thereby are then concatenated and pooled using max-over-time pooling (see Figure 10). Pooling is necessary to transform the word representations that the convolutions produce into a representation for the whole document. We apply dropout with a dropout rate of 0.1, meaning that at each update step, a randomly selected 10% of the weights is neglected. This is performed to prevent the model from learning dependencies between specific features, producing a more robust model. We used Adadelta [53] for the training that is a variant of stochastic gradient descent that uses adaptive learning rates per dimension to improve convergence. The neural network is trained [53] for 30 epochs with a learning rate of 1, followed by 20 epochs of 0.1. Such a learning rate schedule can help with fine-tuning the model weights. Weight decay is applied with a factor of 1×10^{-4} to regularise the model and prevent it from overfitting.

Compared to a baseline model that uses a simple approach of term frequency–inverse document frequency (tf-idf) weighted bags-of-words fed into a linear Support Vector Machine [54,55], it repeatedly produces significantly better results. It achieves an AP of 0.871 ± 0.0048 versus 0.858 ± 0.005 for the baseline model on the test set. Figure 11 shows boxplots with median values for the AP values of the baseline and CNN model.

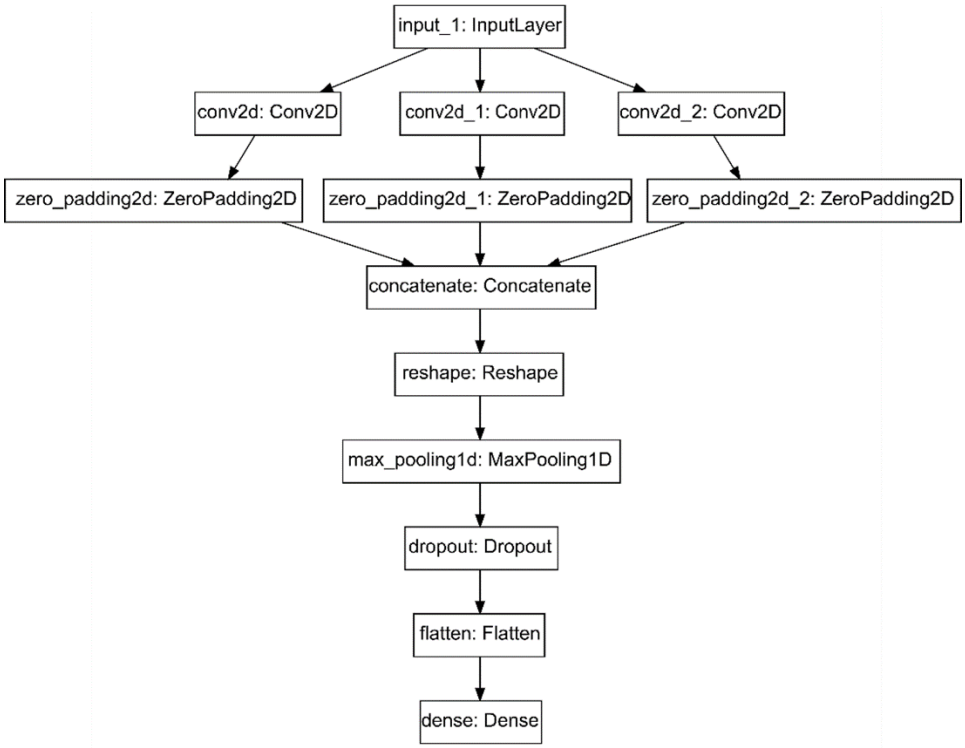

Figure 10. Convolutional Neural Network architecture including flattening layers.

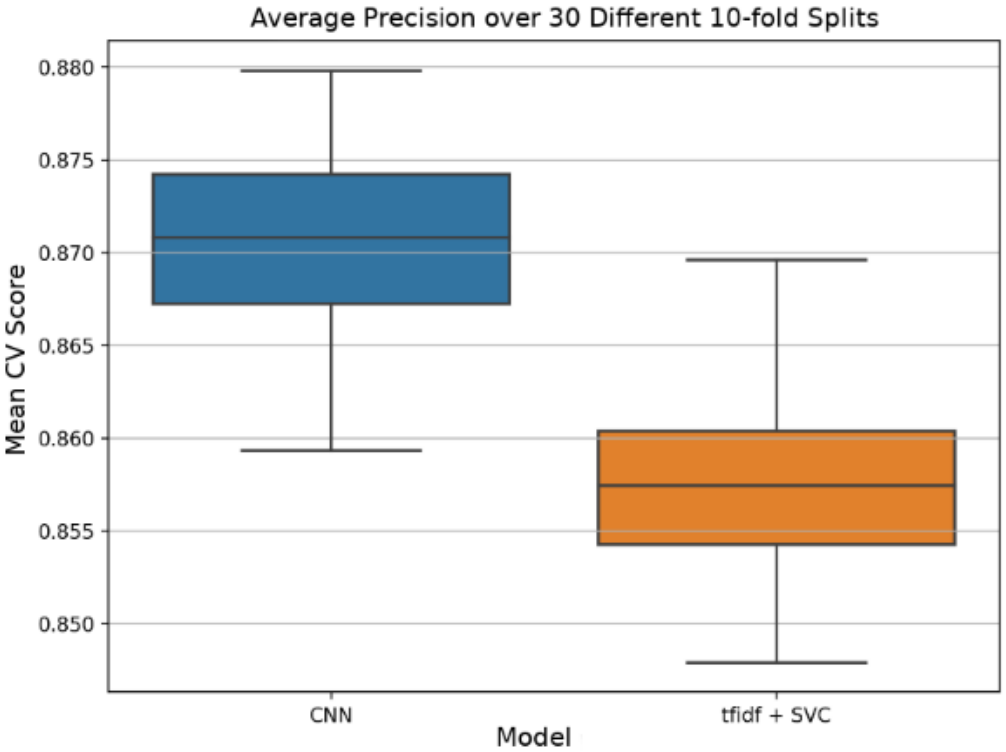

Figure 11. Comparison of Convolutional Neural Network and baseline model.

4.3.2. Dependency Parsing and Location Extraction with DBpedia

We extracted movement information from the text using the dependency parser from StanfordNLP Core [56] to find keywords related to previously detected locations. It is available for multiple languages. An example of such a result is given in Figure 12. Note how the dependency numeric modifier (nummod) can be used to detect which

object a quantity is referring to, with pobj providing information about which objects prepositions refer to. Furthermore, we added artificial "dependencies" via a high-precision heuristic technique in which any sequence of the form "<preposition> <noun>" as well as "<verb> <preposition> <noun>" is considered a direct dependency. This is especially useful for determining movement direction. The keywords such as "refugee" and the prepositions "to" and "from" for the example of Figure 12 need to be manually selected for each language.

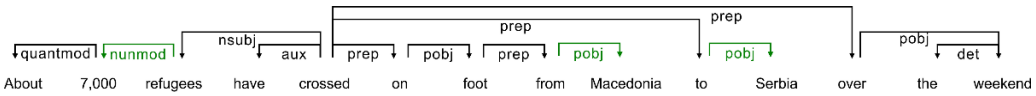

Figure 12. Dependency parsing results for a selected tweet. Highlighted in green are dependencies that indicate cardinality, origin, and destination.

We used DBpedia Spotlight to extract the locations from the tweet texts. DBpedia Spotlight is a tool used to link unstructured information to DBpedia resources. The Wikipedia dataset is the basis of DBpedia, which is structured as an open knowledge graph. DBpedia Spotlight uses the characteristics of the knowledge graph for spotting. Minimum support, defined as 20 Wikipedia in-links and a spotting confidence of 0.5, is required for a match. To avoid ambiguity, i.e., the linking of multiple entities to one surface form, DBpedia Spotlight only returns the best candidate for a phrase, as per the internal confidence score. The Simple Protocol and RDF Query Language (SPARQL) endpoint of DBpedia makes it possible to determine the longitude and latitude for location names that are in DBpedia. Information about a location's area and its type (e.g., continent, country, municipality) is also retrieved via this endpoint.

4.3.3. Deduplication of Tweets and Authenticity Scores

As reposting is frequent in microblogging, textual duplicates need to be removed. This is no trivial task due to the slight changes that may be applied before reposting, such as user mentions or other comments. However, the limited number of characters encountered allows us to use a simplification in this task, namely the binary BoW on n-grams with n = 3. The Jaccard similarity between these bags is then compared against a threshold. Since we do not need all pairwise similarities but only find similar pairs, an inverted index can be used to drastically reduce computation. Each post is only compared against posts that contain at least one common n-gram, retrieved via the inverted index. The relatively high n = 3 ensures that this results in a low number of necessary comparisons.

Authenticity scores should reflect how likely it is that factual information represented in a post is truthful. During the United States presidential election in 2016, "fake news" was shared on social media platforms, which led to awareness strategies to reduce its impact [57,58]. In addition, several social networks reacted and Twitter added an additional label to tweets with synthetic and manipulated media [59]. Based on these efforts, we assume a basic level of user scrutiny before reposting. Accordingly, a post is considered more authentic if it accumulates many duplicates in a corpus of relevant instances [60]. Overall, this means that authenticity scores are a by-product of the above-mentioned textual deduplication. The authenticity score can be weighted by using the Jaccard similarity of the deduplication.

4.4. Forecasting Daily Arrivals of Refugees

A time series is defined by a stochastic process $(y_t;\ t = 1, 2, \ldots, n)$, which represents an ordered series of values with the index time t. The forecasting task involves predicting the unseen values ahead of time. In our analyses, several deterministic and stochastic approaches were applied to the data to estimate the daily arrivals of refugees through the Balkans.

We first produced an autoregressive (AR(p)) model, where p is the order of the model, and the model is given by $y_t = c + \varnothing_1 y_{t-1} + \epsilon_t$, where ϵ_t is white noise, and constant c describes a stationary process whenever the autoregression parameter $\varnothing_1 \in (-1, 1)$. Next, we established an autoregressive–moving-average (ARMA) model with a moving averages (MA(p)) component added to the AR(p) model, given by $y_t = c + \varnothing_1 y_{t-1} + \theta\epsilon_t$. An ARIMA model explicitly assumes the underlying trend to be a linear function. By adding a new constant and an additional time-dependent term to the previous equations, the ARIMA (p = autoregressive, d = difference and q = moving averages) can be expressed as:

$$y_t = c(1 - \varnothing) + b((1 - \varnothing)t + \varnothing) + \varnothing_1 y_{t-1} + \theta\epsilon_t, \tag{2}$$

Next, we applied a seasonal model that models the seasonality by extending the idea of ordered differencing in ARIMA by forming seasonal differences. s is the seasonal period for the differenced series $w_t = z_t - z_{t-s} = \nabla^d \nabla_s^D Y_t$ if it satisfies an $ARMA(p,q)X(P,Q)_s$. Y_t is denoted as $ARIMA(p,d,q) \times (P,D,Q)_s$ and the model can be expressed as:

$$\Phi(B)\phi(B)\nabla_s^D \nabla^d Y_t = \Theta(B)\theta(B)Z_t. \tag{3}$$

The model has sets of parameters, namely, the non-seasonal (p, d and q) and the seasonal parameters (P, D and Q), which are determined in a three-fold process: identification, parameter estimation, and diagnostic validation. We visualised the standardised residuals of a SARIMA model in Figure 13 that was used in the diagnostic validation. For a good fit, the residuals should be normally distributed around zero, which can also be seen in Figure 13. At the identification stage, an ARMA process is developed based on the estimated autocorrelated function (ACF) and partial autocorrelation functions (PACF). The order of the model is chosen based on the lowest Akaike's Information Criterion (AIC) [61]. In the next step, residual monitoring is performed to determine the adequacy of the model, and the process repeats until the model fits certain expected criteria, for instance, the residuals being normally distributed around zero.

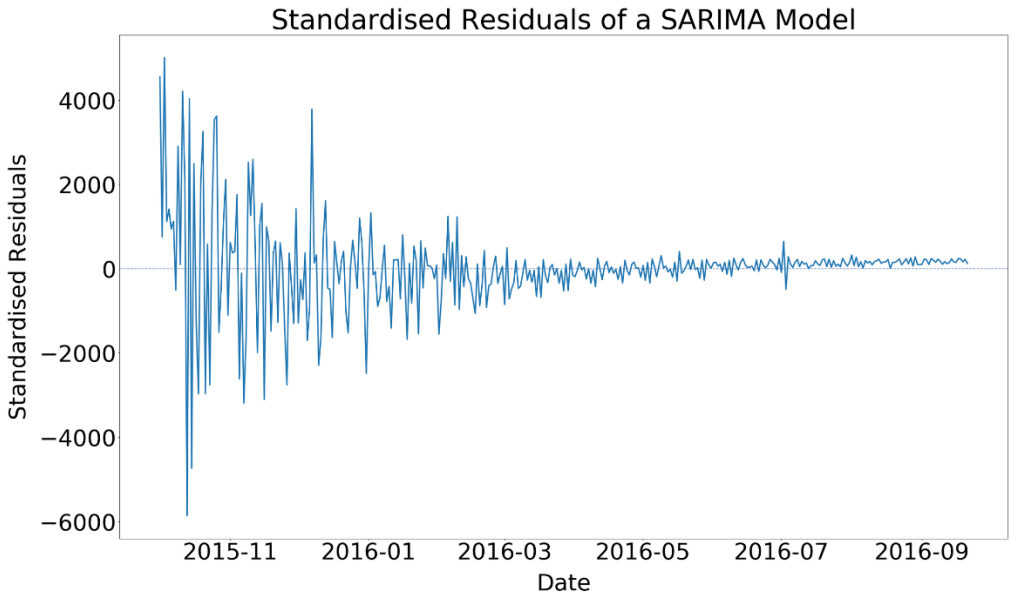

Figure 13. An illustration of the diagnostic validation phase during the fitting of SARIMA models.

Further, we train a simple Bayesian network, which can be expressed as a set of observation equation y_t and the system equation θ_t.

$$\begin{aligned} y_t &= F_t\theta_t + \epsilon_t; \epsilon_t \sim N(0, V_t), \\ \theta_t &= G_t x_{t-1} + w_t; w_t \sim N(0, W_t), \\ \theta_0 &\sim N(m_0, C_0), \end{aligned} \tag{4}$$

θ_0 serves as initial information, and m_0 and C_0 are known p-dimensional and $p \times p$ dimensional vectors. F_t and G_t, also known as the evolution matrix, are known p-dimensional vectors of covariates. V_t and W_t describe the covariate structure among the components of y_t and θ_t, respectively.

A first-order derivative is calculated for the expected value at each step of y_t using the different models to estimate the slope of the trend curve. This trend curve is included to reflect important information regarding the degree of increase or decrease in incoming refugees.

4.5. Designing Network Models for Refugee Movements

Forecasting methods are limited to the availability of data and can seldom be extrapolated to include unobserved situations. This is particularly true for refugee movement, where political decisions influence the model at hand. Thus, we explored geo-spatial simulation models to overcome the limitation of data-driven forecasting methods.

Particularly, we investigated a network-based simulation model to simulate refugee movements. In a network, the vertices of the graph are an abstraction of a geographic reference area such as a state name, a spatial resolution, or coordinates. Its edges denote possible paths between those geographic references. Furthermore, the graph is directed and contains weights reflecting the transition probability from one vertex to another vertex. To remain consistent with probabilities, all outgoing edges of every single vertex of the network must sum up to one and must be greater than zero. The resulting adjacency matrix of this graph then has the property of being column stochastic and the matrix is then called a transition matrix. This is very closely related to concepts from a Markov chain, but the vertices are geographical states rather than events.

A network with fixed transition probabilities behaves like a column stochastic matrix, and when iterating over a time index, it will converge to its eigenvalues, as shown in the example in Figure 14. To take a further temporal dynamic into account, we allowed the underlying transition probabilities of fixed connectivity of the network to change. For discrete-time steps $t \in \{1, \ldots, T\}$, we have transition probabilities in an adjacency matrix G_t in which only the edges of an initial G_0 (the initial network) are allowed.

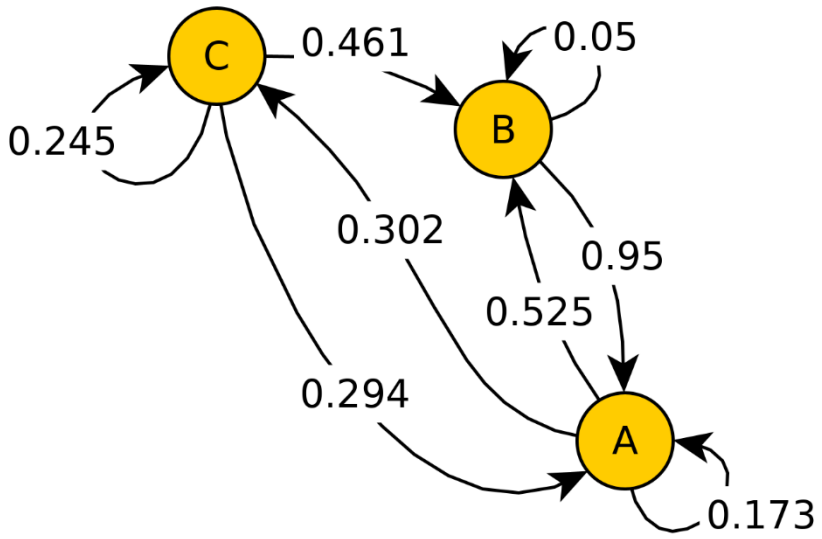

Figure 14. A snapshot of a network with transition probabilities. Single vertices represent geographic areas, and the edges between the paths can be covered within a fixed time window.

To initialise a network, we looked at different stages: defining the network vertices and their geographical mapping, defining initial connectivity, which must be respected in subsequent steps, and defining factors and dynamics, which can change the connectivity in subsequent time steps.

An approach is to manually define a spatial extent in a geographic hierarchy such as the NUTS hierarchy and then come up with semi-automatically generated vertices and their associated geographic regions. Another approach would be to define coordinates for reverse geo-coding and obtain vertices with automatically detected extents of their associated geographic region. We used the NUTS code hierarchy and OpenStreetMap (OSM) for these automated approaches. In the second stage, we collected distance information of the associated geographic regions of vertices, restricted the distances to possible travelling distances for the desired time window, e.g., one day, and used this to define the basic connectivity of the network. In the third stage, we introduced weights (also called factors) for several sources of real-world statistics to control the changing dynamics of the transition probabilities. Those weights can be drawn from a Dirichlet distribution. We consider geographic distances and population densities as two real-world statistics, but the concept easily transfers to other statistics such as weather conditions or air traffic [62].

Finally, the described network can either be used to learn from exemplary time series data or its structure, and the underlying parameterised distributions can be used to sample simulated observations for each region. Given a set of observed regions with statistics about refugees, we can formulate an objective that is then used to improve the network dynamics to sample realistic observations.

5. Results: Spatio-Temporal Results of Refugee Movements

5.1. RQ1: Spatio-Temporal Hot Spot Maps Related to Refugee Movements

As our analyses focus on the refugee movements in 2015, we show exemplary results of our spatio-temporal analysis in the timeframe from February until October 2015 on the Balkan route. At the beginning of 2015, the hot spot maps mostly show hot spots at the border of Turkey and Greece and on the coast of Turkey close to Greece, where refugees gathered to begin their journey. This can be seen in Figure 15. All other areas are considered less relevant, as the number of refugee-related tweets posted in these cells considering the values of neighbouring cells was insignificant. These observations are also in line with official reports on the refugee movements that state that collective refugee movements started in the northern hemisphere in spring 2015 [63].

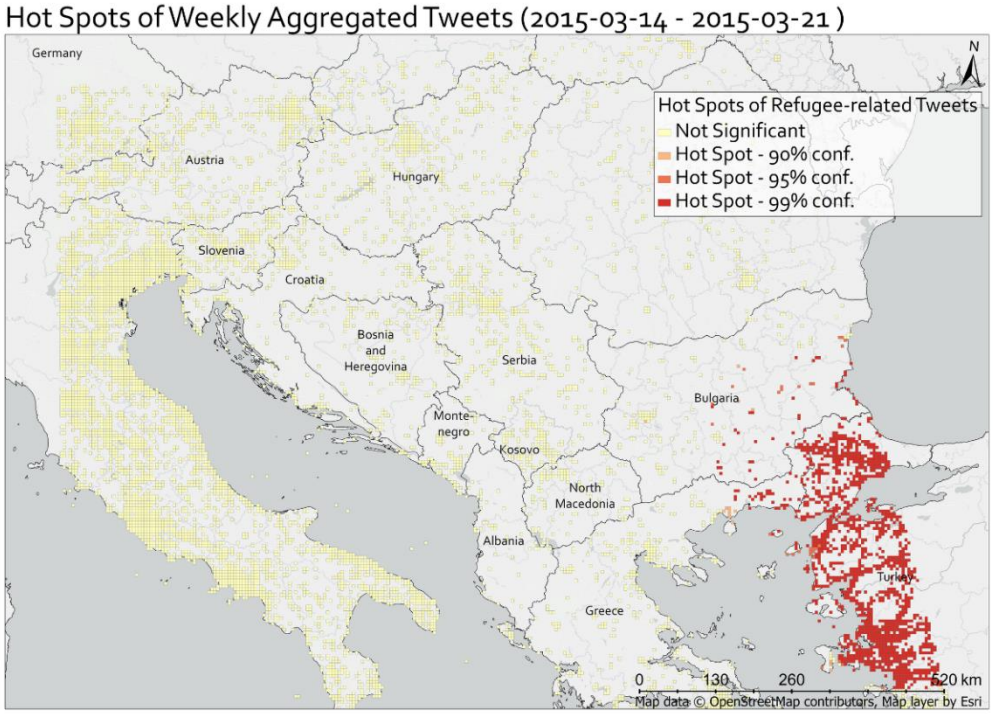

Figure 15. Weekly hot spots of aggregated relevant tweets from 14 March 2015–21 March 2015. Reprinted from map data © OpenStreetMap contributors, map layer by Esri, under a CC BY 4.0 license.

During the autumn of 2015, the hot spots shifted more and more towards the borders of North Macedonia, Serbia, Hungary, Austria and Germany, as can be seen in Figure 16. This route was also the most common path of the refugee movements from Turkey to Germany, demonstrating that the hot spot maps show spatial patterns related to the refugee movements and the primary path.

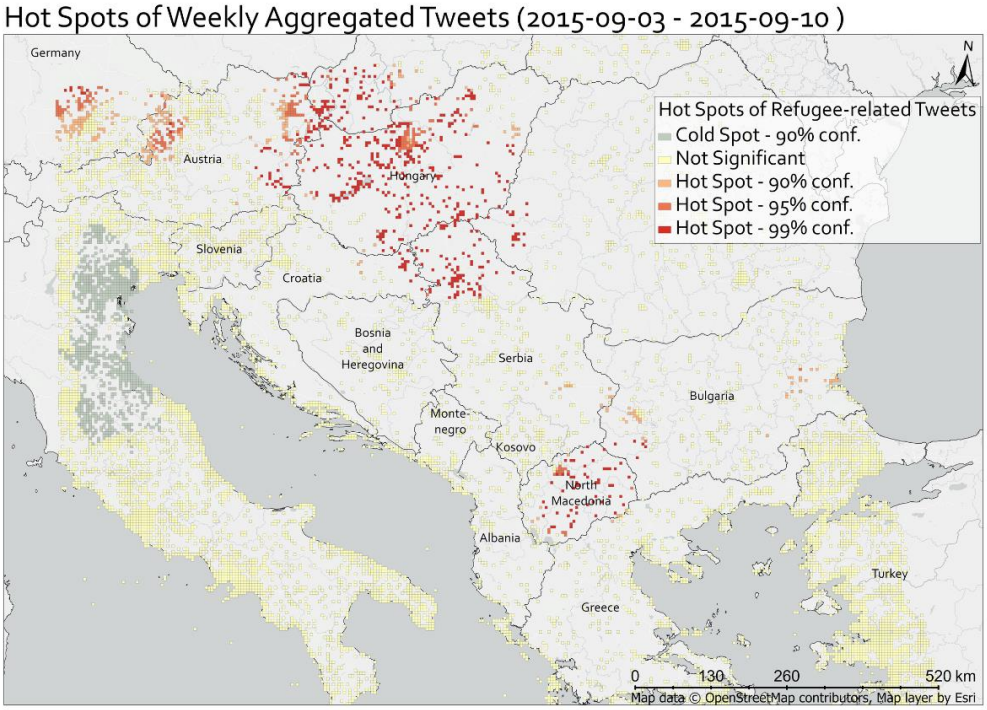

Figure 16. Weekly hot spots of aggregated relevant tweets from 3 September 2015–10 September 2015. Reprinted from map data © OpenStreetMap contributors, map layer by Esri, under a CC BY 4.0 license.

In 2015, the authorities of the countries along the Balkan route all made decisions that were not coordinated with the authorities of the other affected countries. For example, Hungary built a fence along its borders with Serbia and Croatia that drastically reduced the number of refugees travelling through Hungary. In response, refugees made their way through Croatia, Slovenia, and Austria to reach their destinations. This is also visible in the hot spot maps where this change led to hot spots along the new route, as can be seen in Figure 17.

Besides those examples, we observed that the hot spots are either on the route of the refugees to Central Europe or at the western border of Turkey, where most refugees began their journey. By considering examples and our own observations, we conclude that the Twitter data-based hot spot maps reflect the real geo-spatial events related to refugee movements and specifically that the spatially clustered high occurrence of refugee-related social media posts corresponds with the actual refugee movements.

5.2. RQ2: Visualise Factual Information Extracted from Tweets

Using the extracted movement information that is explained in Section 4.3, we created chord diagrams to visualise the movements described in the text of the General-Tweets and the Geo-Tweets. A total of 3914 tweets were classified as refugee-relevant tweets by the CNN, and 1242 tweets were removed through deduplication.

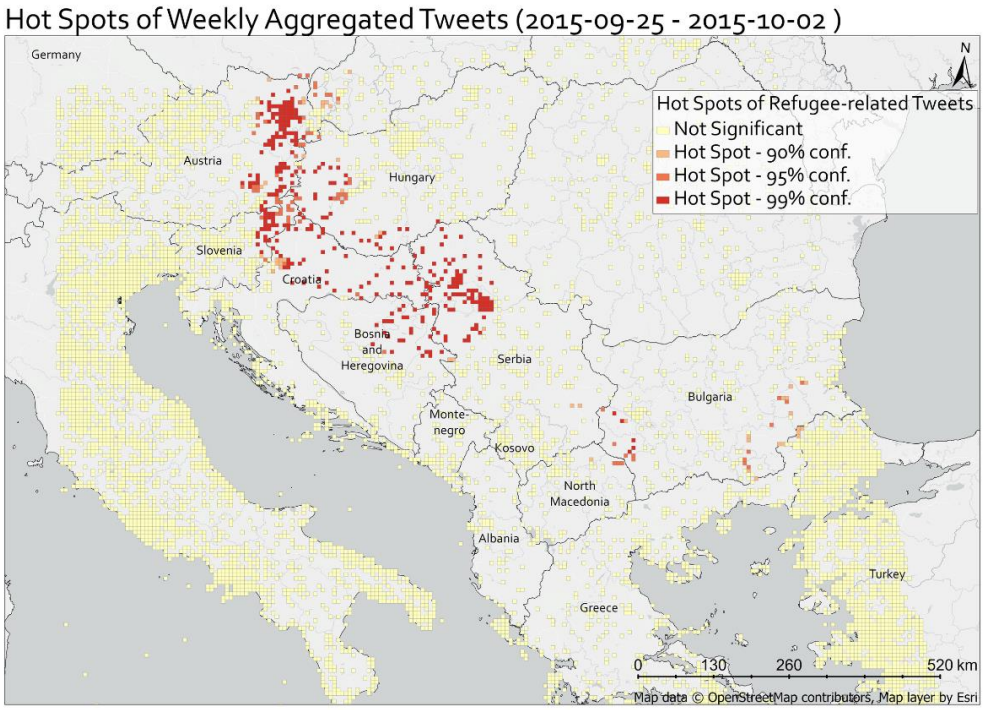

Figure 17. Weekly hot spots of aggregated relevant tweets from 25 September 2015–2 October 2015. Reprinted from map data © OpenStreetMap contributors, map layer by Esri, under a CC BY 4.0 license.

From each tweet text, we identified the mentioned origin, the destination, and the number of refugees. After this information extraction, an aggregation of the number of refugees took place across different origins and destinations. Both the origin and the destination are required to depict a flow and create a chord diagram. So, where no origin or destination could be identified, it was classified as "undefined". The arcs of the diagram are depicted as arrows pointing from the origin to the destination, and the axes show the number of refugees. In Figure 18, the extracted quantities are aggregated, and in Figure 19, the number of tweets is aggregated.

In the chord diagrams, we can see that Hungary is the most important transit country, as large amounts of people name the country as a destination and as a place of origin. Especially Budapest is often mentioned as the origin because thousands of refugees temporarily stopped at Budapest's main train station, and for many refugees, this was the last stop before reaching Germany or Austria, which are the most desired destination countries according to the diagram [64]. Flows through Balkan countries are not detected as well as the flow between Hungary, Austria, and Germany. This might be due to less frequent use of Twitter in these countries. However, some of them are listed in the diagram because they were mentioned in refugee-related tweets.

The mentioned countries in the chord diagrams include Italy, Greece, North Macedonia, Croatia, Serbia, Hungary, Slovenia, and Austria, which are all the countries in the UNHCR dataset. This overlap shows that the identified countries in the chord diagrams do also represent the countries along the Balkan route. However, the numbers of the arrived refugees in the UNHCR dataset should not be compared with the extracted values from social media as social media does not cover all the refugee movements along the route but rather focuses on specific events and observations.

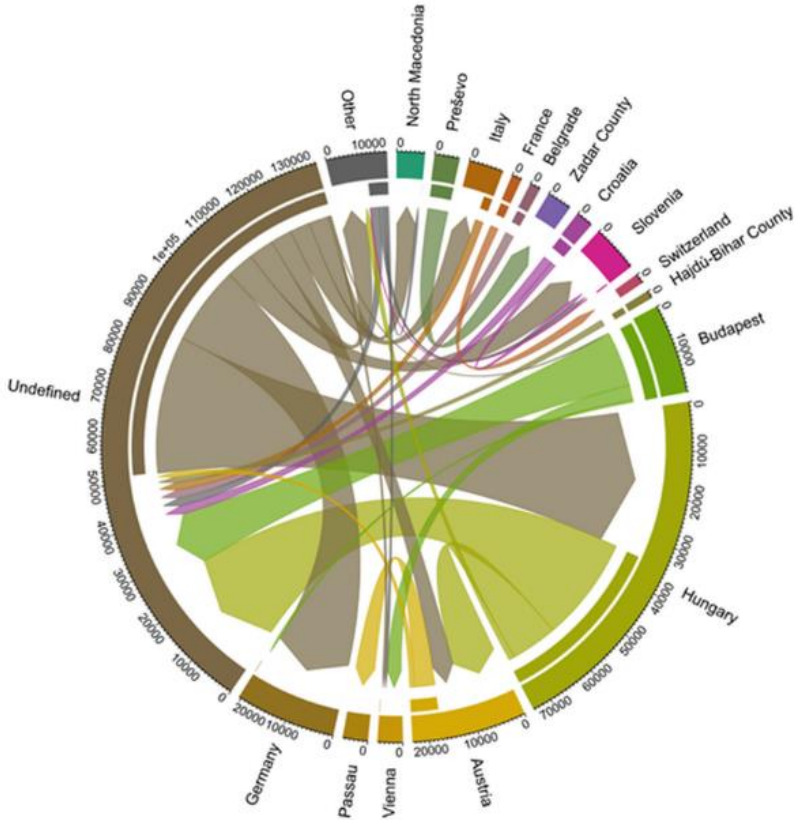

Figure 18. Chord diagram showing the flow of refugees across the Balkan route (based on numbers mentioned in the text).

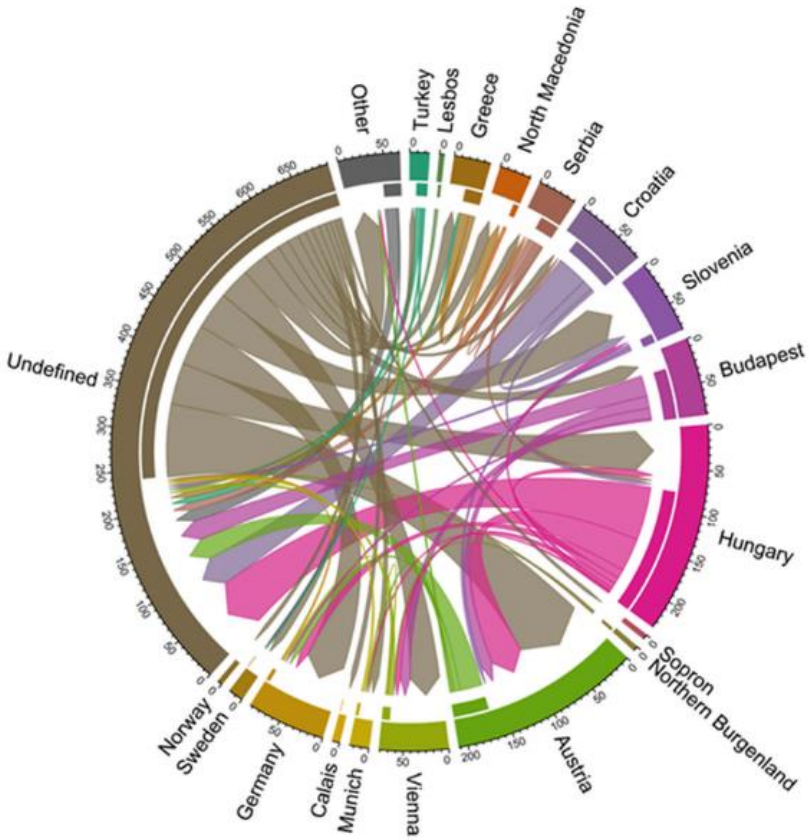

Figure 19. Chord diagram showing the flow of refugees across the Balkan route (number of tweets).

5.3. RQ3: Forecasting Daily Incoming Refugees per Country

We calculated dynamic forecasts with estimates up to four days in advance along with frequentist confidence intervals (CIs). The forecasts consider no observed information about the dependent variable during the period they cover. The CIs give a probabilistic estimate of the real value to lie within the range; in our case, a random selection would result in the true value to lie in the given range at least 95% of the time. The chosen parameters for the seasonal ARIMA were (1,1,1) for both the seasonal and non-seasonal components, based on the observed choice of parameters, as discussed in the relevant literature [36] and the exploratory results in Section 3.2. Figure 20 illustrates the results of the different models used for quantitative forecasts. The models are compared quantitatively with the Root Mean Square Error (RMSE) that measures the difference of the forecasted values and the observed values [65]. Overall, ARIMA has the best average RMSE, followed by AR and SARIMA, indicating a robust autoregressive component.

Model	Italy	Greek Islands	Greece	North Macedonia	Serbia	Croatia	Hungary	Slovenia	Austria
AR(1)	435.90	684.94	815.49	511.02	512.55	503.42	93.94	185.67	447.31
MA(1)	513.26	1805.04	1701.35	1845.88	1751.85	1959.49	911.75	1747.19	1936.50
ARIMA(1,1,1,)	480.65	584.78	684.01	478.92	484.85	465.73	95.83	440.87	408.78
$SARIMA(1,1,1)_{12}$	441.42	599.92	811.015	533.49	583.95	555.65	187.58	501.39	549.56
DLM	732.58	1220.42	1438.91	1284.33	1310.62	1225.31	1023.51	1607.47	1258

Figure 20. Average RMSE on one day ahead dynamic forecasts for daily arrivals with the various forecasting methods.

An example of the ARIMA model is shown in Figure 21, where the forecast of daily arrivals in Austria is presented with CIs and trend slopes for a short period instead of the full period as in Figure 20. However, there is no one single best model for forecasts for all the time series.

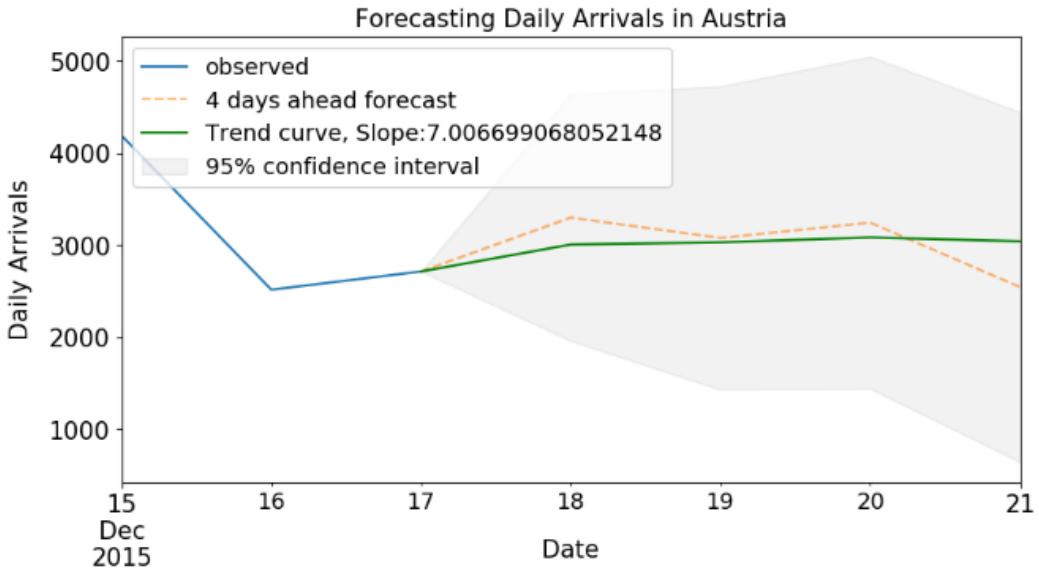

Figure 21. Forecasting daily arrivals in Austria, with confidence intervals and trend slope.

Furthermore, to estimate the accuracy of the trend forecast, each forecast event, regardless of the duration, was treated as an individual data point and, thus, an error rate was calculated. The error rate is 33%, implying there is a two-thirds chance of accurately estimating an increase or decrease in flow.

5.4. RQ4: Network Model for Simulating Refugee Movements

Based on the observed data of the UNHCR, we chose a manual or semi-automatic selection of the NUTS codes hierarchy, as depicted in Figure 22. We obtained a

graph G0 of vertices $V = \{A, B, C\}$. Figure 22 shows a larger practical example with $\{SK, HU, HR, DE1, \ldots, DEG, AT11, AT12, \ldots, SI01, SI02\}$ for this step, where the abbreviations DE, AT, SK, HU and HR stand for Germany, Austria, Slovakia, Hungary and Croatia. The two codes DE and AT are depicted at the state level. Six out of sixteen states for DE and three regions of AT are depicted and connected within their according hierarchy with dashed lines. Directly bordering regions on every level are connected in the graph with solid lines.

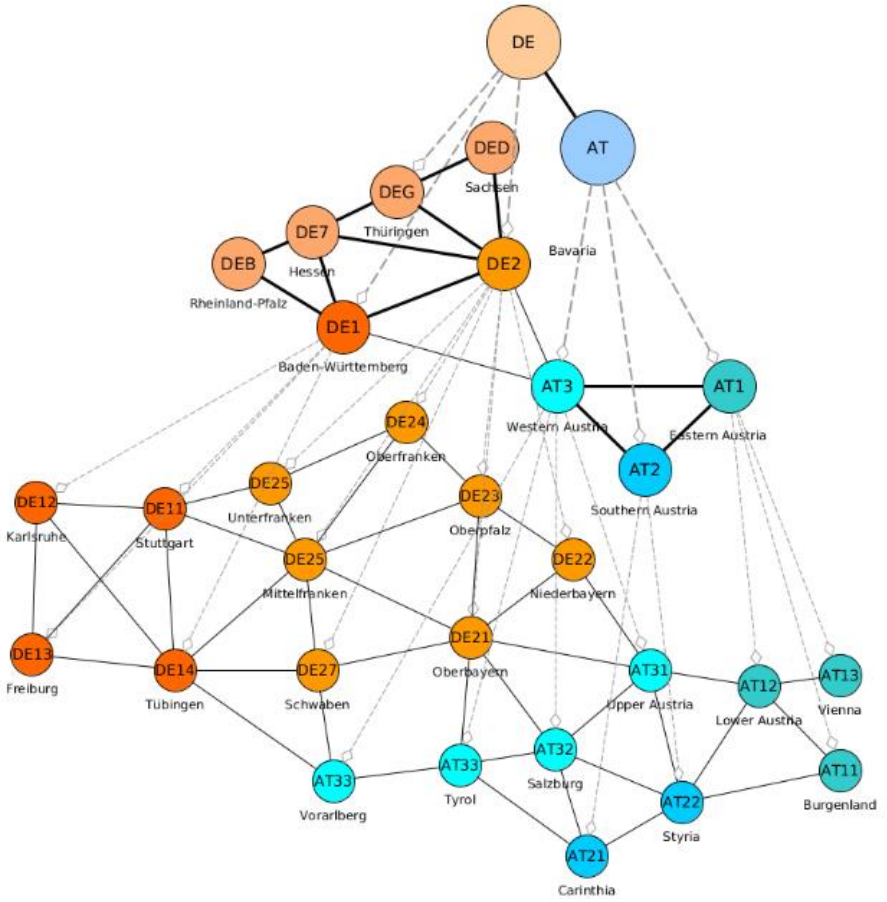

Figure 22. A visualisation of a connected and hierarchical graph of NUTS for some regions of southern Germany and Austria.

OSM and reverse geo-coding were used to calculate the distances between all vertices while a threshold value, such as $\theta d0 = 120$ (kilometres), restricts the connections of the graph. OSM could return a distance matrix for $G0$ given as:

$$D = \begin{pmatrix} 0 & 101 & 146 \\ 101 & 0 & 91 \\ 146 & 91 & 0 \end{pmatrix} \in \mathbb{R}^{|V| \times |V|}, \tag{5}$$

in which $ds,t \in \mathbb{R}$ is the distance between region $s \in V$ and $t \in V$. The complete graph $G0$ is reduced to only connections between geographically close regions by using the threshold value θ, which changes the matrix to:

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \tag{6}$$

Another hyperparameter that needs to be manually defined is the global vector between two geographical points. The angle between this global vector and each edge is used to determine a direction for the edge. That way, we can create a directed graph that represents the Balkan route. An example of the resulting directed graph is shown in Figure 23. As we are interested in simulating or forecasting movements into Austria and Germany, the spatial resolution for these two countries is kept higher than for the neighbouring countries. To validate the simulation, the UNHCR data only provide refugee information at the spatial resolution of countries.

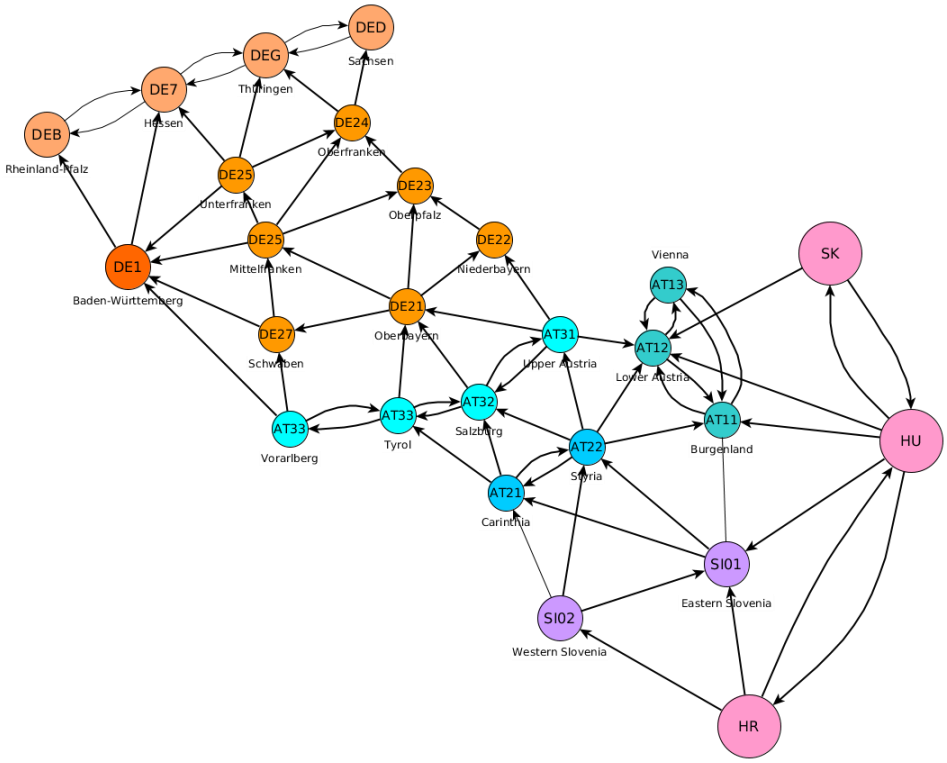

Figure 23. A visualisation of a possible resulting graph obtained from the NUTS hierarchy and distance information of OSM.

Next, we constructed a transition matrix $T \in R^{|V|\times|V|}$ based on the obtained directed network. We obtained a vector $P = \begin{pmatrix} 350 & 100 & 500 \end{pmatrix} \in R^{|V|}$ with $p_v \in P$ being the average population density for region v ∈ V through the NASA Earth Observations [66]. Based on population density (or any other statistics for a region), a self-recurrent transition weight is defined. Aside from the self-recurrent transition weight, other learnable factors, such as the geographical distance or weather conditions, are included that can be drawn from a joint Dirichlet distribution. Finally, we constructed a transition matrix based on a network that is parametrised with learnable factors.

We obtained a simulation model based on a directed network. Given observations of refugees from UNHCR, social media or any hypothetical human sensors, the given network can digest those observed numbers and provide estimates of refugee numbers travelling through these geographic regions. While the model does not explore all possible properties, a corresponding extension should easily be possible.

6. Discussion

We proposed an approach to analyse social media data and refugee movement statistics with state-of-the-art machine learning methods. The experiments cover the refugee movements from Turkey to Central Europe in 2015 and 2016 and show which information about this event can be extracted and how refugee movement statistics can be used with our approach. We evaluated the social media results by comparing them with events and

routes described in newspaper articles. The forecasting and the simulation of the refugee movements rely on the actual refugee movement statistics and the forecast models were assessed quantitatively with the RMSE.

The four research questions defined in the introduction can be answered as follows and comprise the major findings of our analysis:

- RQ1: We found that a hot spot analysis applied on refugee-related tweets extracted using keywords suggested by experts reveals spatio-temporal patterns corresponding to real collective refugee movements described in newspaper articles. The results presented in Section 5.1 demonstrate that changes in the refugee paths as well as key locations were identified in the analysed Twitter dataset.
- RQ2: When training a CNN with annotated data that distinguishes between tweets containing factual information about refugees and others, we were able to identify tweets with information about collective refugee movements (see Section 5.2). The information extracted from social media corresponds to the actual transit routes of refugees at the time.
- RQ3: We applied various time-series models on refugee movement statistics to forecast numbers of arriving refugees per country and found out that, overall, ARIMA achieved the best RMSE (see Section 5.3). However, in some cases, other models, such as AR and SARIMA, outperformed ARIMA.
- RQ4: By combining refugee movement statistics with geo-spatial datasets such as OSM or NUTS, we were able to design a network model to simulate potential refugee movements (see Section 5.4). It simulates state changes per (fixed) time step and estimates how movements occur quantitatively. The transition matrix can be extended with real-world statistics such as population densities, weather conditions or geographical distances that provide a high flexibility for simulation.

The demonstrated methods and datasets provide manifold opportunities to enhance refugee management. First, by extracting and visualising the weekly bins, we can monitor temporal changes within the area of interest where collective refugee movements have been indicated. These changes are of high interest for relief organisations and public authorities who rely on timely movement information to allocate their resources accordingly. Furthermore, the acquired information can be combined with other systems to validate new unverified information provided by third parties. Because our approach allows us to monitor large areas, it can also inform the target selection process for focal monitoring systems such as satellite-based remote sensing.

Secondly, the information extracted from the tweets' content allows us to describe refugee movements quantitatively and qualitatively. This information includes attributes such as the means of transport or the number of refugees travelling from one location to another. By extracting quantitative figures, organisations can optimise their planning and scale their logistics according to the number of refugees arriving.

Thirdly, forecasting based on refugee movement statistics provides organisations with the possibility to go from simply reacting to the situation to being prepared for similar future scenarios ahead of time. In combination with the spatio-temporal simulation, authorities can optimise their logistics based on past events and can simulate potential situations. In comparison to the information extracted from social media, these results are based on statistical datasets that can complement these results well.

Concluding, the research conducted in this study improves the current refugee management procedures by extracting additional information from an online data source and estimating spatial and temporal developments. In comparison to traditional data, the results can be produced in real-time as the social media posts can be collected at any time.

The datasets and methods used in this study provide valuable insights for relief organisations and public authorities. However, multiple aspects need to be considered when using these datasets and methods, which we discuss in this section.

6.1. Limitations and Possible Improvements of Social Media Data Analysis

The Twitter datasets mostly consist of irrelevant data that do not add any value to our analysis. Therefore, we extracted relevant tweets based on predefined keywords. We considered tweets to be relevant if they included one or more keywords. Although we carefully considered the used keywords with the help of experts, the keywords need to be adapted and must be translated by language experts to other languages to make them usable for other or similar use cases in the future. While conducting the analyses to identify relevant tweets, we also tried a topic modelling method to extract language-independent topics from the corpus. Topic modelling has been successfully applied in various other domains to identify relevant tweets. In our case, the goal was to identify one or more topics related to refugee movements. However, topic modelling did not yield a satisfying result. We believe that in comparison to other events, such as earthquakes, for example, refugee movements do not cause a sudden interest in social media and associated activities are distributed over a longer period of time. Therefore, the tweets that are related to refugee movements are not considered to be an independent topic but rather a part of other topics.

The hot spot analysis results showed highly promising results when we compared the hot spot maps with reports of the area. However, one issue with this method is that the calculated cell size strongly influences the result. For the cell size, we adapted a formula that takes the extent of the area and the number of tweets into account. In its original form, we needed to use a larger moving window than weekly so that the cells are fine-grained enough to draw conclusions about areas of particular interest, such as borders. In the case of the weekly moving window, such areas of interest were often only represented by one cell. In order to use the weekly moving window, we adapted the formula and added a variable in the divisor. This makes the use of the formula more flexible for various use cases and domains.

By using only keywords for the extraction of relevant tweets in the first step, we are aware that the extracted dataset can include artificially created tweets by bots or fake news about refugee movements. Both problems need to be tackled, but there is no general solution for detecting bots and fake news yet. Therefore, we decided not to remove any tweets from the dataset and assumed that the number of artificially created tweets and fake tweets does not strongly influence the results.

Another critical aspect is the ambiguities in the location extraction from the text by DBpedia Spotlight, whereby the location with the highest internal score is matched with the location in the text. DBpedia Spotlight detects multiple locations, but the score can still be improved. The internal score depends on the context of the text, which improves with the length of the text. As tweets are short by nature and several studies have shown that DBpedia Spotlight can be worse than other platforms [67], we expect that the results can be improved by combining multiple platforms.

The information extraction could identify either the origin or the destination mentioned in a tweet in most cases, which results in the additional category "undefined" in the chord diagrams. Although this is an unsatisfactory result, it reflects the nature of tweets where people assume a certain knowledge of context and do not share every piece of information explicitly. This could be solved by matching the information from different tweets where either origin or destination is missing.

While the used CNN model achieves a reasonable performance, which is superior compared to a baseline model, it still suffers from a lack of available training data. Data augmentation can be employed to improve the generalisation ability of a neural network. Such strategies apply label-preserving transformation during training and thereby increase the variability of data seen by the network. Zhang et al. [68] propose augmentations by replacing words using a thesaurus. Thereby, words are stochastically replaced with words that have a similar meaning according to the thesaurus. A replacement is chosen from the ranked synonyms via, e.g., a geometric distribution. A similar approach can be undertaken by computing the similarity as the (cosine) distance in the word embedding space. Both

data augmentation strategies were tried but proved to be of little benefit while being costly to compute.

Any strategy that works on the words has the disadvantage of requiring separate preprocessing, including embedding lookups, for each training epoch (e.g., 50×). Furthermore, replacing words via similarity in the embedding domain is costly, even when scaling them so that the Euclidean is equal to the cosine similarity and can be used to increase the computation speed. Lastly, both the distribution for determining whether a word is replaced as well as how a replacement is selected from available candidates need to be tuned.

On a small dataset, such a process is prone to overfitting, leading to a resulting model that is not usable in the real world. Therefore, no augmentation was used for our final model. A technique that is promising and requires little tuning but was not tested is the back-translation augmentation [69]. In this method, documents are automatically translated into another language and back into the original language, resulting in a paraphrased document with similar meaning. Trying such a method can be considered worthwhile in future work, given that an automatic translator that works well on tweets is available.

Instead of artificially increasing the amount of training data, more posts could be annotated using a crowdsourcing platform. As acquiring labelled data was a challenge for our dataset, the annotator instructions and examples should be designed carefully in order to achieve a consistent understanding of the annotation task. Additionally, the relevance of a post could be further distinguished into categories based on the presence of certain factual information, i.e., locations, refugees, or movement. This would allow for a model to better learn the association between certain semantic or syntactic patterns and facts that can be extracted later. If these categories could be learned automatically, an improvement in the runtime and accuracy for fact extraction could be achieved by only performing the steps for the types of information that are indicated by the classifier.

While text-based deduplication is an important step in the information extraction pipeline, it does not guarantee factual consistency. Specifically, it cannot be used to remove different paraphrases of the same information, e.g., headlines of news articles that describe the same event (various tweets in the Twitter datasets describe newspaper articles). If, however, fact extraction works to an appropriate degree, fact-based deduplication can be applied later to either aggregate information between duplicates or remove them. Techniques from both record linking and data deduplication methods needed, e.g., for database systems, could be used in this task.

6.2. Uncertainties in Time-Series Analysis and Reliability of Simulation Results

Refugee movements are inherently a very complex process reliant on a large set of uncertain circumstances. Forecasting such a complex phenomenon is undoubtedly riddled with high quantitative errors, making the choice of a perfect model rather arbitrary. Based on our analyses, it is hard to ascertain which approach is superior solely based on quantitative analysis, and we believe that relying on any one technique would not be prudent. The trend forecast, which is a hybrid approach, yields a more accurate result. Rather than focusing on quantitative single-valued outcomes, the trends in different outcomes should be evaluated subjectively and used in a larger risk management framework comprised of inputs from other experts. Our research demonstrates that a relatively softer mode of estimation, which involves forecasting trends with their corresponding slopes and model-specific uncertainties, can yield valuable insights into risk management frameworks.

The statistical modelling problem of refugee movements across both a temporal and a spatial dimension is not only complex in its mathematical nature but also too highly parameterised for so few time-series data points that we have available from the UNHCR database. This renders our research into a mostly analytical and artificial setting, and while simulation runs could give the impression of realistic results, it is inherently difficult to assess the reliability of their information. Reliability can only be provided if we make sure we analyse multiple settings of refugee movements of rich data quality in both the

temporal and spatial dimensions and can be sure as intelligent humans that the explored settings are of a mostly independent and diverse nature.

From a technical perspective, we see a rising number of tools combining complex network analysis, differential equations, and agent-based simulation models, and our simulation model can serve as a fusion between network analysis and agent-based simulation with close relationships to hidden Markov models. Those tools can advance predictive and simulative systems for, e.g., refugee movements, information spreading, or epidemic modelling.

6.3. Availability of Datasets

The analysed datasets in this study were collected via Twitter's API and UNHCR's website. Social media data are commonly collectable through an API that social media networks such as Twitter, Foursquare, YouTube, or Flickr provide. In general, various parameters (constraining the request to a specific time span, area, or keywords that must be part of the data's text) can be set to target the most useful data for the specific use case. However, the networks also restrict access to their data by limiting the use of their API by setting a maximum number of requests per predefined time slot or only providing data that were posted in the last seven days. Furthermore, it is only possible to collect a small percentage of the full dataset, even when the requirements of the API are met. These limitations can be overcome with paid subscriptions that reduce the restrictions by a network. However, only public data can be collected from the network as social media users can also protect their data by customising their data privacy settings in the social media networks.

Considering the aforementioned possibilities and limitations, public authorities and emergency organisations should set up and deploy a programme to collect social media data before a potential refugee movement and should continuously collect data that they can analyse if a sudden refugee movement occurs.

The UNHCR dataset is freely available on the UNHCR website. The updated estimates for the daily arrivals per country were published irregularly on their website and included the latest updates on previously published data and new data. As the data are provided by one organisation but include information from UNHCR border teams, authorities, and humanitarian partners, data availability strongly depends on UNHCR and their partners. Although there are also other organisations that provide information about refugees, we could not find a dataset with a similar temporal and spatial resolution for this study.

7. Conclusions and Outlook

This paper proposes an approach to analyse social media data and historical refugee movement statistics to extract information about current refugee movements, forecast daily arrivals per country, and simulate refugee movements in a network model. We analysed the temporal, spatial, and semantic features of social media data in a combined approach that advances the status quo in which methods focus mostly on one feature or are limited to experimental results. Our approach provides insights about collective refugee movements by identifying spatial patterns and factual information that correlate with real-world events. The forecast and the simulation based on refugee movement statistics enable us to forecast an increase or decrease in incoming refugees and analyse potential future scenarios.

Our results demonstrate that the analysis of social media data and refugee movement statistics is beneficial for refugee movement. Social media data provide timely information about collective refugee movements that can be new information or complement others. Furthermore, the extracted information from text can be used to describe the refugee movements quantitatively and qualitatively so that relief organisations and public authorities can plan their logistics according to the number of arriving refugees. The forecasting and simulation of refugee movements give relief organisations and public authorities the possibility to not only react to certain situations but also be prepared for future scenarios. However, in future research studies the reliability of the simulation model should be as-

sessed with denser spatio-temporal data and the RMSE of the forecasting models should be improved before including them in monitoring systems.

There are still various unresolved research challenges. The location extraction from the social media posts' content was not validated in this study and should be compared with other meta-data from the social media posts to assess and increase the performance of the location extraction algorithm. The detection of bots and fake news could outline the influence of these methods and would improve the risk assessment of relying on the extracted information. Additionally, the forecasting and simulation results rely on statistical data that only reflect incoming people and not how many currently reside in the country. A dataset about the current number of refugees could strongly decrease the uncertainties and increase the reliability of the results. Lastly, the results of the two data sources should be fused to effectively use them in one system.

Author Contributions: Conceptualization, Clemens Havas, Lorenz Wendlinger, Julian Stier, Sahib Julka, Veronika Krieger, Cornelia Ferner, Michael Granitzer, and Bernd Resch; methodology, Clemens Havas, Lorenz Wendlinger, Julian Stier, Sahib Julka, Veronika Krieger, Michael Granitzer, and Bernd Resch; software, Clemens Havas, Lorenz Wendlinger, Julian Stier, Sahib Julka, Veronika Krieger, Cornelia Ferner, and Andreas Petutschnig; validation, Clemens Havas, Lorenz Wendlinger, Julian Stier, Sahib Julka, Veronika Krieger, Cornelia Ferner, and Andreas Petutschnig; formal analysis, Clemens Havas, Lorenz Wendlinger, Julian Stier, Sahib Julka, Veronika Krieger, Cornelia Ferner, and Andreas Petutschnig; investigation, Clemens Havas, Lorenz Wendlinger, Julian Stier, Sahib Julka, and Veronika Krieger; resources, Michael Granitzer and Bernd Resch; data curation, Clemens Havas, Lorenz Wendlinger, Julian Stier, Sahib Julka, and Veronika Krieger; writing—original draft preparation, Clemens Havas, Lorenz Wendlinger, Julian Stier, Sahib Julka, Veronika Krieger, Cornelia Ferner, Andreas Petutschnig, Michael Granitzer, Stefan Wegenkittl, and Bernd Resch; writing—review and editing, Clemens Havas; visualization, Clemens Havas, Lorenz Wendlinger, Julian Stier, Sahib Julka, Veronika Krieger, Cornelia Ferner, and Andreas Petutschnig; supervision, Michael Granitzer, Stefan Wegenkittl, and Bernd Resch; project administration, Michael Granitzer and Bernd Resch; funding acquisition, Michael Granitzer and Bernd Resch. All authors have read and agreed to the published version of the manuscript.

Funding: This study was carried out as part of the HUMAN+ project, which is funded by the Austrian security research programme KIRAS of the Federal Ministry of Agriculture, Regions and Tourism (BMLRT), project number 865697. It was also supported by the Austrian Science Fund (FWF) through the Doctoral College GIScience at the University of Salzburg (DK W 1237-N23).

Data Availability Statement: The Geo-Tweets dataset is available via Harvard Dataverse. Please follow the link to access the dataset: https://doi.org/10.7910/DVN/VS6COJ (accessed on 21 July 2021). The other datasets were collected from public repositories that are referenced in the manuscript.

Acknowledgments: We would like to thank Harvard University's Center for Geographic Analysis for their support in providing us with the Twitter data for our study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Eurostat. Asylum and First Time Asylum Applicants-Annual Aggregated Data (Rounded) [Internet]. Available online: https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tps00191&plugin=1 (accessed on 12 March 2020).
2. El-Shaarawi, N.; Razsa, M. Movements upon movements: Refugee and activist struggles to open the Balkan route to Europe. *Hist. Anthropol.* **2019**, *30*, 91–112. [CrossRef]
3. Weber, J. Migrationsdruck durch Flüchtlinge: Die Südostbayerischen Grenzräume am Ende der Balkanroute 2015–2016. In *Grenzüberschreitende Raumentwicklung Bayerns: Dynamik in der Kooperation-Potenziale der Verflechtung*; Verlag der ARL—Akademie für Raumforschung und Landesplanung: Hannover, Germany, 2018; pp. 159–186.
4. Kostakos, V.; Rogstadius, J.; Ferreira, D.; Hosio, S.; Goncalves, J. Human sensors. In *Participatory Sensing, Opinions and Collective Awareness*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 69–92.
5. Brunwasser, M. A 21st-century migrant's essentials: Food, shelter, smartphone. *The New York Times*, 26 August 2015.
6. Resch, B. People as sensors and collective sensing-contextual observations complementing geo-sensor network measurements. In *Progress in Location-Based Services*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 391–406.
7. Ostrand, N. The Syrian refugee crisis: A comparison of responses by Germany, Sweden, the United Kingdom, and the United States. *J. Migr. Hum. Secur.* **2015**, *3*, 255–279. [CrossRef]

8. Carrera, S.; Blockmans, S.; Gros, D.; Guild, E. The EU's response to the refugee crisis: Taking stock and setting policy priorities. *CEPS Essay* **2015**, *20*, 1–24.
9. Greussing, E.; Boomgaarden, H.G. Shifting the refugee narrative? An automated frame analysis of Europe's 2015 refugee crisis. *J. Ethn. Migr. Stud.* **2017**, *43*, 1749–1774. [CrossRef]
10. Guiraudon, V. The 2015 refugee crisis was not a turning point: Explaining policy inertia in EU border control. *Eur. Polit Sci.* **2018**, *17*, 151–160. [CrossRef]
11. Gillespie, M.; Osseiran, S.; Cheesman, M. Syrian refugees and the digital passage to Europe: Smartphone infrastructures and affordances. *Soc. Media Soc.* **2018**, *4*, 2056305118764440. [CrossRef]
12. Dekker, R.; Engbersen, G.; Klaver, J.; Vonk, H. Smart refugees: How Syrian asylum migrants use social media information in migration decision-making. *Soc. Media+ Soc.* **2018**, *4*, 2056305118764439. [CrossRef]
13. Curry, T.; Croitoru, A.; Crooks, A.; Stefanidis, A. Exodus 2.0: Crowdsourcing geographical and social trails of mass migration. *J. Geogr. Syst.* **2019**, *21*, 161–187. [CrossRef]
14. Hübl, F.; Cvetojevic, S.; Hochmair, H.; Paulus, G. Analyzing refugee migration patterns using geo-tagged tweets. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 302. [CrossRef]
15. Petutschnig, A.; Havas, C.; Resch, B.; Krieger, V.; Ferner, C. Exploratory Spatiotemporal Language Analysis of Geo-Social Network Data for Identifying Movements of Refugees. *GI_Forum* **2019**, *7*, 137–152.
16. Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, arXiv:180301271.
17. Kim, Y. Convolutional neural networks for sentence classification. *arXiv* **2014**, arXiv:14085882.
18. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A convolutional neural network for modelling sentences. *arXiv* **2014**, arXiv:14042188.
19. Johnson, R.; Zhang, T. Effective use of word order for text categorization with convolutional neural networks. *arXiv* **2014**, arXiv:14121058.
20. Johnson, R.; Zhang, T. Deep pyramid convolutional neural networks for text categorization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 562–570.
21. Aleskerov, F.; Meshcheryakova, N.; Rezyapova, A.; Shvydun, S. Network analysis of international migration. In Proceedings of the International Conference on Network Analysis, Nizhny Novgorod, Russia, 26–28 May 2016; pp. 177–185.
22. Liu, W.; Hou, Q.; Xie, Z.; Mai, X. Urban Network and Regions in China: An Analysis of Daily Migration with Complex Networks Model. *Sustainability* **2020**, *12*, 3208. [CrossRef]
23. De Carvalho, R.C.; Charles-Edwards, E. The evolution of spatial networks of migration in Brazil between 1980 and 2010. *Popul. Space Place* **2020**, *26*, e2332. [CrossRef]
24. Danchev, V.; Porter, M.A. Neither global nor local: Heterogeneous connectivity in spatial network structures of world migration. *Soc. Netw.* **2018**, *53*, 4–19. [CrossRef]
25. Lin, L.; Carley, K.M.; Cheng, S.-F. An agent-based approach to human migration movement. In Proceedings of the 2016 Winter Simulation Conference (WSC), Arlington, VA, USA, 11–14 December 2016; pp. 3510–3520.
26. Suleimenova, D.; Bell, D.; Groen, D. A generalized simulation development approach for predicting refugee destinations. *Sci. Rep.* **2017**, *7*, 1–13. [CrossRef] [PubMed]
27. Rossetti, G.; Milli, L.; Rinzivillo, S.; Sirbu, A.; Pedreschi, D.; Giannotti, F. Ndlib: Studying network diffusion dynamics. In Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 19–21 October 2017; pp. 155–164.
28. Donges, J.F.; Heitzig, J.; Beronov, B.; Wiedermann, M.; Runge, J.; Feng, Q.Y.; Tupikina, L.; Stolbova, V.; Donner, R.V.; Marwan, N.; et al. Unified functional network and nonlinear time series analysis for complex systems science: The pyunicorn package. *Chaos Interdiscip. J. Nonlinear Sci.* **2015**, *25*, 113101. [CrossRef] [PubMed]
29. Bijak, J. *Forecasting International Migration in Europe: A Bayesian View*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2010; Volume 24.
30. Saboia, J.L.M. Autoregressive integrated moving average (ARIMA) models for birth forecasting. *J. Am. Stat. Assoc.* **1977**, *72*, 264–270. [CrossRef]
31. Bijak, J. Forecasting international migration: Selected theories, models, and methods. In Proceedings of the Central European Forum For Migration Research, Warsaw, Poland, April 2006. Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.125.1745&rep=rep1&type=pdf (accessed on 20 July 2021).
32. Nichiforov, C.; Stamatescu, I.; Făgărăşan, I.; Stamatescu, G. Energy consumption forecasting using ARIMA and neural network models. In Proceedings of the 2017 5th International Symposium on Electrical and Electronics Engineering (ISEEE), Galati, Romania, 20–22 October 2017; pp. 1–4.
33. Lee, R.D. Probabilistic approaches to population forecasting. *Popul. Dev. Rev.* **1998**, *24*, 156–190. [CrossRef]
34. Intriligator, M.D.; Bodkin, R.G.; Hsiao, C. *Econometric Models, Techniques, and Applications*; Prentice Hall International Inc.: Upper Saddle River, NJ, USA, 1996.
35. Cohen, J.E.; Roig, M.; Reuman, D.C.; GoGwilt, C. International migration beyond gravity: A statistical model for use in population projections. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 15269–15274. [CrossRef]

36. Bijak, J.; Disney, G.; Findlay, A.M.; Forster, J.J.; Smith, P.W.F.; Wiśniowski, A. Assessing time series models for forecasting international migration: Lessons from the United Kingdom. *J. Forecast.* **2019**, *38*, 470–487. [CrossRef]
37. Lewis, B. Harvard CGA Geotweet Archive v2.0 [Internet]. V2 ed. Harvard Dataverse. 2016. Available online: https://doi.org/10.7910/DVN/3NCMB6 (accessed on 21 July 2021).
38. Wang, Y.; Callan, J.; Zheng, B. Should we use the sample? Analyzing datasets sampled from Twitter's stream API. *ACM Trans. Web* **2015**, *9*, 1–23. [CrossRef]
39. Scott, J. Archive Team: The Twitter Stream Grab [Internet]. 2012. Available online: https://archive.org/details/twitterstream (accessed on 16 April 2021).
40. Urchs, S.; Wendlinger, L.; Mitrović, J.; Granitzer, M. MMoveT15: A Twitter Dataset for Extracting and Analysing Migration-Movement Data of the European Migration Crisis 2015. In Proceedings of the 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Capri, Italy, 12–14 June 2019; pp. 146–149.
41. UNHCR. Daily Estimated Arrivals through Western Balkans Route [Internet]. 2021. Available online: https://data.humdata.org/dataset/daily-estimated-arrivals-through-western-balkans-route (accessed on 15 March 2021).
42. Shaheen, K. Isis "Controls 50% of Syria" after Seizing Historic City of Palmyra [Internet]. 2015. Available online: https://www.theguardian.com/world/2015/may/21/isis-palmyra-syria-islamic-state (accessed on 21 August 2020).
43. Fahim, K.; Bernard, A. Russia Makes an Impact in Syrian Battle for Control of Aleppo [Internet]. 2015. Available online: https://www.nytimes.com/2015/10/21/world/middleeast/russia-makes-an-impact-in-syrian-battle-for-control-of-aleppo.html (accessed on 21 August 2020).
44. Google. Google Trends [Internet]. 2020. Available online: https://support.google.com/trends/answer/6248105?hl=en-GB&ref_topic=6248052 (accessed on 21 August 2020).
45. Google. FAQ about Google Trends Data [Internet]. 2020. Available online: https://support.google.com/trends/answer/4365533?hl=en (accessed on 14 September 2020).
46. Barisione, M.; Michailidou, A.; Airoldi, M. Understanding a digital movement of opinion: The case of #RefugeesWelcome. *Inf. Commun. Soc.* **2019**, *22*, 1145–1164.
47. Ord, J.K.; Getis, A. Local spatial autocorrelation statistics: Distributional issues and an application. *Geogr. Anal.* **1995**, *27*, 286–306. [CrossRef]
48. Wong, D.W.-S.; Lee, J. *Statistical Analysis of Geographic Information with ArcView GIS and ArcGIS*; John Wiley & Sons: Hoboken, NJ, USA, 2005.
49. Daiber, J.; Jakob, M.; Hokamp, C.; Mendes, P.N. Improving efficiency and accuracy in multilingual entity extraction. In Proceedings of the 9th International Conference on Semantic Systems, Graz, Austria, 4–6 September 2013; pp. 121–124.
50. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 722–735.
51. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
52. Henderson, P.; Ferrari, V. End-to-end training of object class detectors for mean average precision. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 198–213.
53. Zeiler, M.D. Adadelta: An adaptive learning rate method. *arXiv* **2012**, arXiv:12125701.
54. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
55. Reyes-Menendez, A.; Saura, J.R.; Filipe, F. Marketing challenges in the# MeToo era: Gaining business insights using an exploratory sentiment analysis. *Heliyon* **2020**, *6*, e03626. [PubMed]
56. Manning, C.D.; Surdeanu, M.; Bauer, J.; Finkel, J.R.; Bethard, S.; McClosky, D. The stanford corenlp natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MD, USA, 23-24 June 2014; pp. 55–60.
57. Apuke, O.D.; Omar, B. Fake news proliferation in Nigeria: Consequences, motivations, and prevention through awareness strategies. *Humanit. Soc. Sci. Rev.* **2020**, *8*, 318–327.
58. Bovet, A.; Makse, H.A. Influence of fake news in Twitter during the 2016 US presidential election. *Nat. Commun.* **2019**, *10*, 1–14. [CrossRef] [PubMed]
59. Roth, Y.; Pickles, N. Updating Our Approach to Misleading Information [Internet]. 2020. Available online: https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html (accessed on 15 March 2020).
60. Yang, Q.; Tufts, C.; Ungar, L.; Guntuku, S.; Merchant, R. To retweet or not to retweet: Understanding what features of cardiovascular tweets influence their retransmission. *J. Health Commun.* **2018**, *23*, 1026–1035. [CrossRef] [PubMed]
61. Akaik, H. Information theory and an extension of the maximum likelihood principle. In Proceedings of the Second International Symposium on Information Theory, Tsahkadsor, Armenia, 2–8 September 1971; Akademiai Kiado: Budapest, Hungary, 1973; pp. 267–281.
62. Hsu, D.A. Detecting shifts of parameter in gamma sequences with applications to stock price and air traffic flow analysis. *J. Am. Stat. Assoc.* **1979**, *74*, 31–40. [CrossRef]
63. Heisbourg, F. The strategic implications of the Syrian refugee crisis. *Survival* **2015**, *57*, 7–20. [CrossRef]

64. Murray, D. Europe's Growing Refugee and Migration Crisis on Show in Hungary [Internet]. 2015. Available online: https://www.unhcr.org/news/latest/2015/9/55e9dd346/europes-growing-refugee-migration-crisis-show-hungary.html (accessed on 4 September 2020).
65. Chujai, P.; Kerdprasop, N.; Kerdprasop, K. Time series analysis of household electric consumption with ARIMA and ARMA models. In Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong, China, 13–15 March 2013; pp. 295–300.
66. Center for International Earth Science Information Network—CIESIN—Columbia University. *Gridded Population of the World, Version 4 (GPWv4): Population Density*; NASA Socioeconomic Data and Applications Center (SEDAC): Palisades, NY, USA, 2016.
67. Rizzo, G.; Troncy, R.; Hellmann, S.; Bruemmer, M. NERD meets NIF: Lifting NLP Extraction Results to the Linked Data Cloud. In Proceedings of the 5th International Workshop on Linked Data on the Web, Heraklion, Greece, 27 May 2012; p. 937.
68. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 649–657.
69. Sennrich, R.; Haddow, B.; Birch, A. Improving neural machine translation models with monolingual data. *arXiv* **2015**, arXiv:151106709.