Predictive Analytics for Population Growth of Refugees in Asia

Umar Ibn Ali & Md. Shaikot Hossen
Department of Computer Science, KICT, IIUM
umarcs1994@gmail.com

Abstract

The refugee population worldwide is on the rise. Refugees are leaving their origins to migrate to places they believe can give them basic humanitarian needs and in many cases a much-anticipated escape from conflict. The countries that are receiving these refugees are often ill-prepared for the huge surplus in the population of refugees in the recent past. This creates a situation where these refugees are unattended and the governments unprepared. Predicting the population growth of Refugees in a country is a complex task since there are a lot of factors involved in the migration process. In this research project, we build a model to predict the number of refugees in each Asian country which has an average population of 2000 refugees in the last 26 years and an average population of 2000 refugees in the last decade. The model for each country takes several factors which affect the refugee population, as input and predicts the number of refugees for that country up-to 2022 as output. The tweets of the people of the Asian countries are analyzed from twitter. The tweets are classified into positive, negative and neutral sentimentsl.

1. Introduction

Predictive analytics provides the ability to extract meaningful information from vast amounts of data allowing us to identify patterns and trends, make connections between seemingly unrelated data sets, and predict future outcomes [1]. We have used this approach to model population growth of refugees in Asian countries.

When it comes to modelling, there are several ways to look at the current refugee influx in Asia in the recent years. We can analyze their origins and the conflict involved that made them leave their countries. This process involves prediction of violence and military intervention of the hosting countries. We do not use their origin information since this is difficult to predict with the data that is available right now.

Another method for prediction is modelling based on population growth of refugees in hosting countries. This method accounts for the refugees in the hosting countries but does not account for the origins of these refugee population. Gorlach and Motz (2017) mention that in many contexts of armed conflict, most displaced individuals only have the immediate option to escape to a nearby country [2]. If the population of these hosting countries is analyzed, we can model to predict the population growth of refugees in these countries and in turn also the population growth of refugees in Asia.

In this project, separate predictive models are being built for each Asian country which have a refugee concern. Countries with refugee population mean of less than 2000 in the last 26 years were excluded, as the net refugee population there can be thought to be negligible. Countries with refugee population mean of less than 2000 in the last decade were also excluded, as this indicates that these countries do not have a refugee concern in the recent past. Countries which are mainly known for producing refugees were also excluded, for example Syria. Following the above analogy, there were 20 countries for which the predictive model of population growth of refugees was built.

For the second part of this project, it is important to analyze the attitude of ordinary citizens of each Asian country towards refugees. To gather the data, Twitter's large, diverse dataset is used that reflects the public's live opinions and emotions [3]. We collected tweets for every Asian country with at least 5 tweets about refugees per year, from 2014 to 2017. The tweets that used English language were collected only.

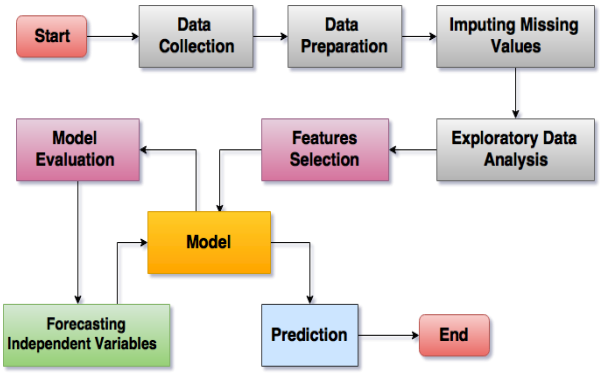
2. Literature Review

UNHCR estimates that there are approximately 60 million people displaced worldwide with 42,500 newly displaced each day. With the ability to predict trends in migration patterns and understand the behavior of people escaping conflict, the efficiency of governments and organizations working to support displaced populations could be improved. [1] Research published in Nature showed that the number of people migrating between two places is highly determined by the distance involved, the size of population living between the two points, and the socioeconomic level of the people migrating. [4] Dynamic behavioral models have been used to examine internal and international migration by Kirdar (2012) and Llull (2017) [5] [6].

The Swedish Migration Board uses big data and analytics for several years to gain visibility into immigration trends and what those trends will mean for the country. [7]

Pak and Paroubek (2010) proposed a model to classify the tweets as objective, positive and negative. They created a twitter corpus by collecting tweets using Twitter API and automatically annotating those tweets using emoticons [8]. Bing Liu (2012) confirmed that there's a correlation between sentiment measures computed utilizing word frequencies in tweets and both patron self-assurance polls and political polls. [9]

3. Methodology



[Figure 1: Methodology Diagram]

3.1 Datasets and Data Collection

UNHCR Refugee Population Data – UNHCR (United Nations High Commission for Refugees) yearly refugee population data from 1991 to 2016 was collected for each of the countries. UNHCR publishes yearly Population Statistics on their website updated up-to 2016. We used the data labelled "Refugees including refugee like situations". This is our target/dependent variable. [10]

Number of Refugee Countries- From the UNHCR Refugee Population data mentioned above, number of refugee countries dataset from 1991 to 2016 was created by entering data manually. For a given year in the hosting country, it describes the number of origin countries the refugee population is made up of. This is one of our independent variables.

GDP Per Capita- Per capita GDP is a measure of the total output of a country that takes gross domestic product (GDP) and divides it by the number of people in the country. The dataset for GDP per capita was extracted from year 1991 to 2016 from International Monetary Fund (IMF)'s World Economic Outlook (WEO) [11]. This is one of our /independent variables.

Population Density- Population density is the number of people living per unit of an area. It is measured in population per square kilometers. The dataset from year 1991 to 2016 was extracted from The International Monetary Fund (IMF)'s World Economic Outlook (WEO). This is another independent variable.

Global Peace Index (GPI)- Global Peace Index (GPI) ranks 163 independent states and territories from 2008 to 2016 according to their level of peacefulness, produced by the Institute for Economics and Peace (IEP) [12]. GPI investigates 23 different factors regarding current domestic and international conflicts of a country and provides a score which determines the peace ranking. The lower the score, the more peaceful a country is. The missing values of the dataset is imputed with mean. This is another independent variable.

Twitter Data- TweetDeck is a social media dashboard application for management of Twitter accounts. The dashboard contains a search option. The search can be made by starting and ending date, location, keyword and language. Tweets were manually collected for every country in Asia with at least 5 tweets matching keyword 'refugee' per year from 2014 to 2017.

3.2 Data Preparation

The UNHCR Refugee Population, Number of Refugee country, GDP per capita, GPI and Population Density were used as columns to make the dataset of each country. The dataset had yearly values from 1991 to 2016. For each country, there are 26 rows and 6 columns of data. A portion of Malaysia's dataset is shown in Table 1. The model is being built for 20 countries, thus 20 datasets needed to be prepared.

Year	GPI	GDP	No of country	Density	Population
2016	1.648	9374	48	94.4	92287
2015	1.561	9505	49	93.5	94166
2014	1.659	11009	44	92	99381
2013	1.574	10700	41	90.4	97513
2012	1.59	10655	43	88.33	90185
2011	1.485	10253	39	86.97	86667

[Table 1: Portion of Malaysia's Dataset]

For sentiment analysis, tweets collected from TweetDeck were stored in a separate text file for each country. Table 2 demonstrates this arrangement of tweets.

Year	Tweets
2014	Most #Syrian refugee children miss formal education in Turkey
2014	How other countries are responding to the UN's refugee requests
2014	Little hope for Burmese Rohingya in refugee camps @JamilaHanan @Aungaungsittwe
2014	Photograph of a #Refugee within 2 miles of his homeland @Refugees

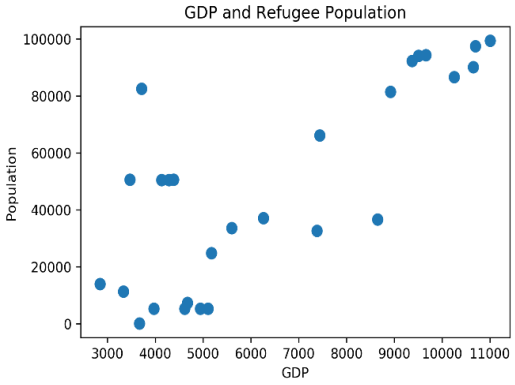
[Table 2: Tweets about Refugees from Bangladesh]

3.3 Imputation of Missing values

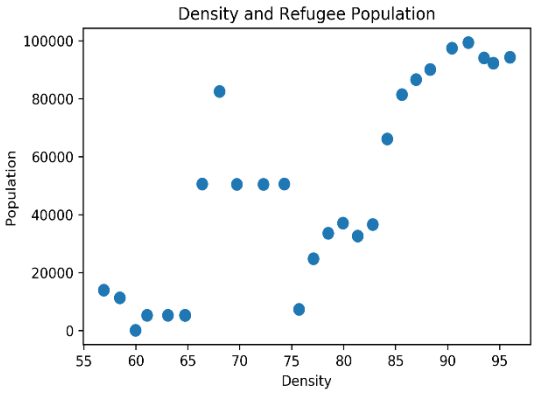
Only the GPI dataset had missing values. The imputation of missing values for each country's GPI dataset was done with mean. This process was done after loading the dataset in pandas data frame.

3.4 Exploratory Data Analysis (EDA)

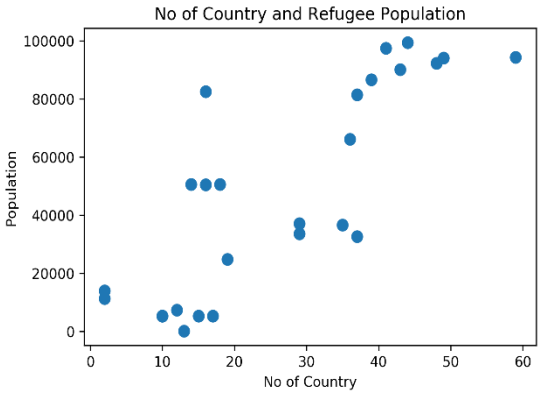
Initially the relationship between the dependent variable (Population of Refugees) and the independent variables (GDP, GPI, Population Density and No of Refugee Country) is analyzed. In the case of Malaysia, there appears to be a good linear relationship between the Population of Refugees and GDP Per Capita, Density and No of Country as shown in figures 2, 3 and 4 respectively.



[Figure 2]



[Figure 3]



[Figure 4]

3.5 Feature Selection

Before prediction and modelling with all the variables, features which may have the strongest correlation with Population of refugees were tested. A Linear Regression model with the dependent variable (Population) was built for each of our independent variables.

3.6 Modelling and Evaluation

Multiple Linear Regression (MLR) was used to build model for prediction of Refugee Population as:

$$\hat{y} = aX1 + bX\ 2 + cX\ 3 + dX\ 4 + \varepsilon \quad (1)$$

The left-hand side of the equation (1) is the predicted value for the population of Refugees in a country for a given year, with X1, X2, X3 and X4 being the predictor values GDP, No of Country and Population Density for that year. The constants a, b, c, d and e are the regression parameters computed by the method of least squares.

There were two scenarios of train-test to avoid overfitting of the model, training 67% and testing 33% and training 75% and testing 25%. Different random

states were also used in scikit-learn to take different sets as training and testing each time.

R squared value was used to check the linear relationship, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values were used to check the prediction accuracy of the independent variables with regards to the dependent variable.

Although, the features were selected in the section above, modelling was done with all different sets of variables. The model was tested each time taking different variables and checking prediction accuracy with R squared, RMSE and MAE values. The model with the highest R square value and lowest RMSE and MAE values was used to do the prediction.

3.7 Forecasting Independent Variables

The next step is to forecast the variables used to build the model. A linear model of each of these variables with the Year was built. After fitting the variables with the year, we predicted the 2017 to 2022 values for each of the variables.

In this step, if the relationship of any of these variables with Year gave a very low R squared score, the model building is repeated by changing the variables.

3.8 Prediction of Refugee Population

The forecasted independent variables were used in the Multiple Linear Regression (MLR) model, made in section 3.6, to predict the Population of Refugees for years 2017 to 2022. The results of the model will be discussed later.

3.9 Sentiment Analysis

TextBlob was used in the python script to analyze the sentiment of the tweets. Polarity score of the tweet was used to categorize the tweets into positive, negative and neutral sentiments. Polarity score of zero indicated that the tweet was neutral, score of more than zero indicated that the tweet was positive, and a negative score indicated that the tweet was negative. The percentage of positive, negative and neutral tweets were calculated and recorded for every country for years 2014 to 2018 using this methodology. The number of tweets for every country in the same range of years were also recorded. The results of Twitter will be displayed under the Results section.

4. Results

4.1 Model Results

The models for each country predicted values up-to the year 2022. Since this is forecasting, 2022 refugee population values cannot be checked by comparison with real values. Thus, the values were checked for 2016 prediction. The table below shows 5 scenarios of predicted 2016 values by the model and actual 2016 values.

Country	Actual Values	Predicted Values
Malaysia	92287	91934
China	317260	308638
Yemen	269796	243703
Tajikistan	2731	2297
Thailand	106471	114070

[Table 3: Actual vs Predicted for 2016]

The table below shows the top 4 countries with the highest predicted population of refugees in 2022. It can be seen from the table, that different countries have different rates of increase of refugee population.

Country	2016	2018	2020	2022
Turkey	1688008	1874707	2061407	2248107
Pakistan	1567471	1581124	1594778	1608432
Lebanon	703453	784967	866481	947995
Jordan	603137	664238	725339	786440

[Table 4: Highest Population in 2022]

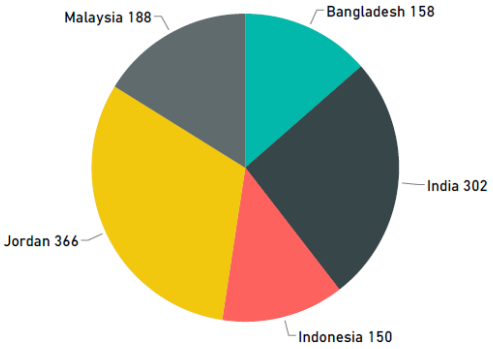
Percentage Increase was calculated by the increase in refugee population from 2016 to 2022. It is to be noted that 2016 predicted values were used to do the calculation. Table below shows the 3 countries with highest predicted increase in population.

Country	2016	2022	Percentage Increase
Lebanon	703453	947995	34.76
Turkey	1688008	2248107	33.18
Jordan	603137	786440	30.39

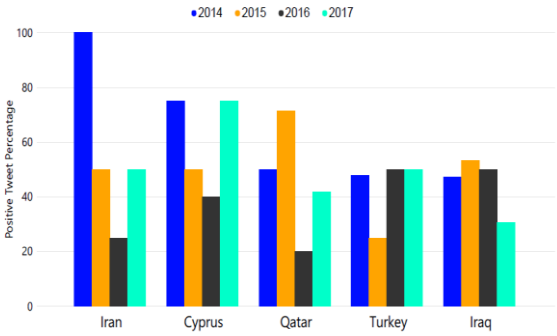
[Table 5: Highest percentage increase]

4.2 Sentiment Analysis Results

The most number of tweets in the years 2014 to 2017 about refugees came from these 5 countries shown in Figure 4. Jordan had the highest number among all the Asian countries.

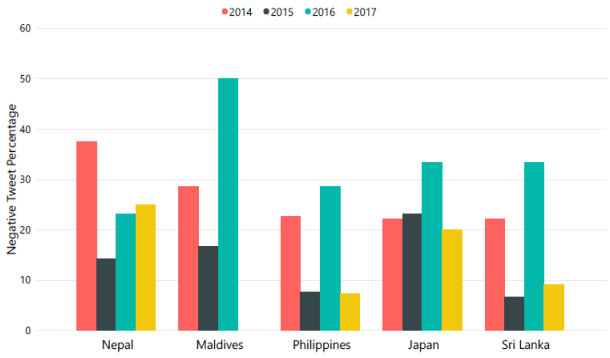


[Figure 4: Countries with most tweets about Refugees]



[Figure 5: Countries with highest percentage of positive tweets]

To calculate the highest number of positive and negative tweet percentage, we found the mean percentage of positive and negative tweets over the course of 4 years. The yearly percentages for the top 5 countries with the highest percentage of positive and negative tweets are shown in figures 5 and 6 respectively.



[Figure 6: Countries with highest percentage of negative tweets]

5. Conclusion

The model predictions showed satisfactory results overall. The 2016 predicted values for some of the models were close to the actual values. A good correlation can be drawn with highest positive percentage of tweets and high predicted refugee population. Turkey was among the top 5 countries with highest percentages of positive tweets and Turkey also had the highest predicted 2022 population. A correlation can also be drawn with the highest number of tweets and predicted refugee population. Jordan and Bangladesh were on the list for most tweets and highest predicted 2022 refugee population.

6. Future Enhancements

6.1 Improvement of Model

There were instances where the relationship between the independent variables and the population of refugees did not follow a linear pattern, primarily because yearly refugee population values could not detect the fluctuations due to ever changing war situations. There were also instances where the forecasting of the independent values with year did not fetch good independent value predictions, in those instances the predicted refugee population did not show great results. The results of the prediction could be greatly improved by monthly refugee population data. The movement of refugees do not follow a pattern, monthly values could detect the fluctuation in the refugee population better than yearly values. Office of UNHCR Malaysia was contacted through email for the monthly values, but there was no response. The GPI is calculated with 23 indicators, including “Number of refugees and displaced persons as percentage of

population", "Financial contribution to UN peacekeeping missions", "Military expenditure as a percentage of GDP" and "Political terror" among others. The erratic increase of refugee population due to war situations of could be better predicted by adding these indicators as independent variables to the existing country datasets. The indicators of GPI mentioned above are not made open to the public.

6.2 Tweet collection using Twitter Search API

Initially keys generated from Twitter's developer account were used to write a python script to get tweets using the keyword 'refugee'. Then the tweets were parsed by location and arranged for performing sentiment analysis. Very few tweets had location data and twitter only allowed accessing and downloading of tweets that were tweeted not more than a week ago. This meant that there was no historical data available by this method. Thus, this approach was aborted, and alternative methods were looked into. TweetDeck was used as the other alternatives such as Gnip are not free of costs.

7. References

[1] Novack, R. (2015, September 22) We Should Have Seen This Refugee Crisis Coming. Retrieved from: www.wired.com/2015/09/able-predict-refugee-crisis/

[2] Gorlach, Motz, (2017) "Refuge and Refugee Migration: How Much of a Pull Factor Are Recognition Rates?"

[3] Brouckman, L, Wang, A., "Analyzing Twitter Sentiment on the Refugee Crisis in 2016", Stanford, 2016

[4] Locker, M (2016, December 9) Predicting The Break: How Nations Can Get Ahead Of The Next Refugee Crisis. Retrieved from: https://www.fastcompany.com/3063078/predicting-the-break-how-nations-can-get-ahead-of-the-next-refugee-crisis

[5] Kirdar, Murat G., "Estimating the Impact of Immigrants on the Host Country Social Security System when Return Migration is an Endogenous Choice," International Economic Review, 2012, 53 (2), 453-486.

[6] Llull, Joan, "Immigration, Wages, and Education: A Labor Market Equilibrium Structural Model," Review of Economic Studies, 2017, 1, 1-46.

[7] Pratt, M.K (2016, February 8) Big data's big role in humanitarian aid. Retrieved from: www.computerworld.com/article/3027117/big data/bigdatas- big-role-in-humanitarian-aid.html

[8] A. Pak and P. Paroubek., Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326

[9] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

[10] UNHCR Population Statistics (2017). Time Series, Popstats.unhcr.org. Adapted from: http://popstats.unhcr.org/en/time_series

[11] World Economic Outlook Database April 2017. (n.d.) Retrieved from https://www.imf.org/external/pubs/ft/weo/2017/01/weodata/index.aspx

[12] Vision of humanity. (2017). Global Peace Index – Vision of Humanity. Retrieved from http://visionofhumanity.org/indexes/global-peace-index