الجامعة الإسلامية العالمية ماليزيا
INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA
يونيفرسيتي إسلام انتارابغسا ملسيا

KULLIYYAH OF INFORMATION AND COMMUNICATION TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

FINAL YEAR PROJECT REPORT

Predictive Analytics for Population Growth of Refugees in Asia

UMAR IBN ALI
1328435

MD. SHAIKOT HOSSEN
1233903

SUPERVISED BY

MADAM SURIANI BT. SULAIMAN

DECEMBER 2017

SEMESTER I, 2017/2018

FINAL YEAR PROJECT REPORT

Predictive Analytics for Population Growth of Refugees in Asia

by

Umar Ibn Ali

1328435

Md. Shaikot Hossen

1233903

Supervised by

Madam Suriani Bt. Sulaiman

In partial fulfilment of the requirement for the Bachelor of Computer Science

Department of Computer Science

Kulliyyah of Information and Communication Technology

International Islamic University Malaysia

December 2017

Semester 1, 2017/2018

Predictive Analytics for Population Growth of Refugees in Asia

by

Umar Ibn Ali
1328435

Md. Shaikot Hossen
1233903

A project paper submitted to the
Department of Computer Science

Kulliyyah of Information and Communication Technology
in partial fulfillment of the requirement for the
Bachelor of Computer Science

Approved by the Examining Committee:

_____

Madam Suriani Bt. Sulaiman, Project Supervisor

International Islamic University Malaysia

December 2017

Semester 1, 2017/2018

CERTIFICATION OF THE ORGINALITY

This is to certify that we are responsible for the work submitted in this project, that the original work is our own except as specified in the references and acknowledgements, and that the original work contained herein have not been taken or done by unspecified sources or persons.

_____

Umar Ibn Ali
(1328435)

_____

Md. Shaikot Hossen
(1233903)

ACKNOWLEDGEMENT

First, we would like to thank Almighty Allah (SWT), for blessing us with the capability to successfully complete this project.

We wish to express our deepest appreciation to our supervisor Madam Suriani Bt. Sulaiman, for her relentless guidance, helpful suggestion, close supervision and moral encouragement to complete this task.

Special thanks to Dr. Hamwira Yaacob, our final year project coordinator and all those who provided us with any kind of support in completing this report.

We would also like to thank our parents for guiding us and constantly encouraging us towards achieving our goals.

Lastly, we sincerely thank all of those who have directly or indirectly helped us in completing this project.

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 sentiments.

Table of Contents

CHAPTER	TITLE	PAGE
1	INTRODUCTION	1
	1.0 Project Overview	1
	1.1 Problem Description	1
	1.2 Project Objective	2
	1.3 Project Scope	2
	1.4 Constraints	2
	1.5 Significance of the Project	3
	1.6 Organization of the Report	3
2	LITERATURE REVIEW	4
	2.1 Predictive Analytics	4
	2.2 Prediction on Migration Patterns	4
	2.3 Using Twitter for Sentiment Analysis	5
3	METHODOLOGY	6
	3.0 Introduction and Method for inclusion	6
	3.1 Datasets and Data Collection	7
	3.1.1 UNHCR Refugee Population Dataset	7
	3.1.2 GDP Per Capita Dataset	9
	3.1.3 Population Density	10
	3.1.4 Number of Refugee Countries	11
	3.1.5 Global Peace Index (GPI)	11
	3.1.6 Twitter Data	12
	3.2 Data Preparation	12

3.3 Imputation of Missing values……. 15
3.4 Exploratory Data Analysis…… 15
3.5 Feature Selection…… 17
3.6 Modelling and Evaluation…… 18
3.7 Forecasting Independent Variables…… 20
3.8 Prediction of Refugee Population…… 21
3.9 Sentiment Analysis…… 22

4 RESULTS ……. 23
4.1 Model Results…… 23
4.1.1 Evaluation of Prediction…… 23
4.1.2 Prediction up to 2022…… 24
4.2 Analysis of Prediction Results…… 25
4.3 Sentiment Analysis Results…… 26

5 CONCLUSION ……. 28
5.1 Correlation……. 28
5.2 Future Enhancement …… 28
5.2.1 Improvement of Model……. 28
5.2.2 Alternate Prediction Technique…… 29
5.3 Conclusion…… 29

REFERENCES 30

CHAPTER 1

INTRODUCTION

1.0 Project Overview

This project is about predicting the number of refugees in Asian countries up-to year 2022 using predictive analytics. The refugee crisis has reached a critical point, with scores of people being displaced every day. This research provides new literature on using predictive analytics for examination and prediction of migration of refugee population around the world. In this project, sentiments of citizens of Asian countries were also analyzed using Twitter data to examine their view on the refugee scenario.

1.1 Problem Description

The refugee population is on the rise. UNHCR (United Nations High Commission for Refugees) estimates that there are approximately 60 million people displaced worldwide with 42,500 newly displaced each day. The countries that are receiving these refugees are not equipped for the huge surplus in the number of refugees in the recent past. This creates a disastrous scenario where human rights are neglected; the NGOs fail to deliver proper shelters or proper refugee camps and the hosting governments are overburdened. Predicting the population of Refugees would mean that the governments of refugee hosting countries and NGOs can prepare for refugee migration beforehand, this would result in better infrastructure and opportunities for the refugees expected to enter a country.

The opinion of the public where refugees are entering is not considered and thus not known. It is important to analyze the tweets of ordinary citizens of countries about refugees to gain insight on their opinions and evaluate if there is a correlation between public opinion and predicted refugee population.

1.2 Project Objective

The project objectives are as follows:

1. Predict the population of refugees in hosting Asian countries up-to year 2022
2. Analyze tweets to evaluate the acceptability of Refugees in these Asian countries.
3. Examine the results of model Prediction and sentiment analysis.

1.3 Project Scope

This project has point of interest to a broad range of audience including Governments, NGOS, Data Scientists, personals working to mitigate the refugee crisis around the globe and Researchers working to examine migration patterns using predictive modelling.

1.4 Constraints

The primary constraint in this project is lack of available datasets. Some datasets are not made available to the public, for example from organizations such as Amnesty, which could have greatly improved the results. Organizations such as UNHCR, were contacted for more detailed datasets, but there was no response.

1.5 Significance of the Project

The significance of this research is huge, even a little advancement in this task would open many doors throughout the world. Governments can fund experts to predict migration trends into their country. The efficiency of organizations such as UNHCR working to support displaced populations can be improved and strategic policy decisions can be made that impact millions of people. Enhancing this idea, population of refugees being produced from a country can be predicted and steps can be taken beforehand to mitigate the cause of refugee production.

1.6 Organization of Report

This project report is organized as follow:

Chapter 1 is an introductory chapter which consists of a brief explanation about our research project on Predictive Analytics for Population Growth of Refugees in Asia.

Chapter 2 is the literature review.

Chapter 3 is the methodology and explanation of steps taken for model prediction and sentiment analysis.

Chapter 4 is the chapter on results of our models and sentiment analysis.

Chapter 5 concludes the project with suggested future enhancement.

CHAPTER 2

LITERATURE REVIEW

2.1 Predictive Analytics

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]. Many different organizations in different sectors use predictive analytics to get insight on data and make important business decisions for the future. It is now being used in public health, to predict trends and personalize treatment of patients. The Oil and Gas, E-commerce and Weather Forecasting sectors are key users of predictive analytics.

2.2 Prediction of Migration Patterns

Migration of population most times follow a pattern. 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. [2]

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] A thorough analysis of all the motivational factors of refugees to flee their homeland can be done, divided into cultural differences, religious differences, different ethnicity, war situations and natural disasters. Dynamic behavioral models have been used to examine internal and international migration by Kirdar (2012) and Llull (2017) [3] [4].

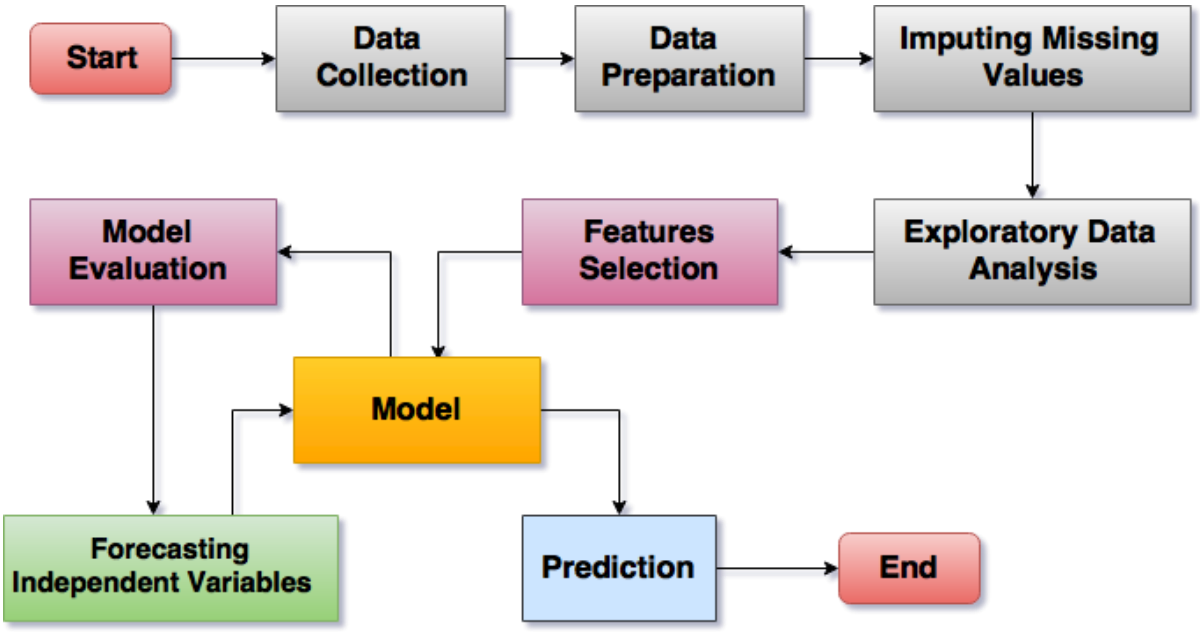
In the recent past, the Swedish Migration Board is using big data and analytics for several years to gain visibility into immigration trends and what those trends will mean for the country. [5] Although, prediction of Refugee population has not been carried out in this scale before, the above instances show that this can be done.

2.3 Using Twitter for Sentiment Analysis

Twitter is one of the most popular social media site in the world. Twitter's large, diverse dataset reflects the public's live opinions and emotions [6]. 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 [7]. There's a growing literature on using twitter to classify sentiments. 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 [8].

CHAPTER 3

METHODOLOGY



[Figure 1: Methodology Diagram for Model Prediction]

3.0 Introduction and Method for inclusion

In this project, separate predictive models were built for each Asian hosting country which have a refugee concern. Countries with yearly refugee population mean of 2000 both in the last 26 years and last 10 years were included. The former indicates that the country does have a refugee concern and the latter indicates that the country does have a refugee concern in the recent past. Countries which are mainly known for producing refugees were also excluded, for example Syria. Using the above methodology, 20 countries were included and thus 20 models were built.

For sentiment analysis part, tweets were collected for every country in Asia with at least 5 tweets matching keyword ‘refugee’ per year from 2014 to 2017.

3.1 Datasets and Data Collection

3.1.1 UNHCR Refugee Population Dataset

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. The data labelled “Refugees including refugee like situations” were used. This is the target/dependent variable [9]. The raw dataset is shown in the table.

Year	Country of asylum/residence	Origin	Population type	Value
1991	Turkey	Iran (Islamic Rep. of)	Refugees (incl. refugee-like situations)	1363
1991	Turkey	Iraq	Refugees (incl. refugee-like situations)	28049
1991	Turkey	Various/Unknown	Refugees (incl. refugee-like situations)	1050
1992	Turkey	Bosnia and Herzegovina	Refugees (incl. refugee-like situations)	15083
1992	Turkey	Iran (Islamic Rep. of)	Refugees (incl. refugee-like situations)	1761
1992	Turkey	Iraq	Refugees (incl. refugee-like situations)	11437
1992	Turkey	Various/Unknown	Refugees (incl. refugee-like situations)	196
1993	Turkey	Bosnia and Herzegovina	Refugees (incl. refugee-like situations)	16773
1993	Turkey	Iran (Islamic Rep. of)	Refugees (incl. refugee-like situations)	1386
1993	Turkey	Iraq	Refugees (incl. refugee-like situations)	4944
1993	Turkey	Various/Unknown	Refugees (incl. refugee-like situations)	164
1994	Turkey	Bosnia and Herzegovina	Refugees (incl. refugee-like situations)	20000
1994	Turkey	Iran (Islamic Rep. of)	Refugees (incl. refugee-like situations)	2008
1994	Turkey	Iraq	Refugees (incl. refugee-like situations)	2672
1994	Turkey	Various/Unknown	Refugees (incl. refugee-like situations)	247
1995	Turkey	Afghanistan	Refugees (incl. refugee-like situations)	1550
1995	Turkey	Bulgaria	Refugees (incl. refugee-like situations)	67
1995	Turkey	Bosnia and Herzegovina	Refugees (incl. refugee-like situations)	5614
1995	Turkey	Greece	Refugees (incl. refugee-like situations)	92

[Table 01: Portion of UNHCR Refugee Population Dataset of Turkey]

3.1.2 GDP Per Capita Dataset

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) [10]. This is one of the independent variables.

Asia	2001	2002	2003	2004	2005	2006	2007	2008	2009
Afghanistan	196	197	197	214	248	270	325	381	435
Armenia	692	779	923	1180	1628	2128	3079	3913	2912
Azerbaijan	665	748	864	1019	1541	2415	3759	5213	4933
Bahrain	13894	13502	14486	15964	17996	19,268	20,908	23,236	19,465
Bangladesh	406	420	454	486	495	523	585	656	728
Bhutan	783	843	933	1056	1172	1288	1502	1865	1721
Brunei	18680	18846	20822	24295	28589	33254	34811	42679	31287
Cambodia	320	338	361	406	471	536	628	742	735
China	1053	1150	1293	1513	1766	2111	2703	3467	3838
Cyprus	14904	16188	20389	23970	25527	27449	31775	36030	32636
Georgia	731	777	924	1202	1522	1863	2479	3159	2694
Hong Kong	25171	24733	23858	24876	26554	28031	30497	31488	30594
India	472	492	572	657	747	837	1077	1049	1153
Indonesia	834	1003	1187	1281	1404	1765	2064	2418	2465
Iran	5081	1951	2269	2671	3202	3733	4758	5435	5419
Israel	20308	18437	18972	19900	20567	21834	24903	29541	27722
Japan	33860	32301	34845	37701	37228	35465	35342	39454	41041
Jordan	1803	1880	1949	2133	2300	2689	2990	2754	3983

[Table 02: Portion of GDP Per Capita Dataset]

3.1.3 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.

Asia	1991	1992	1993	1994	1995	1996	1997	1998	1999
Afghanistan	19.59	21.05	22.71	24.31	25.69	26.78	27.62	28.35	29.16
Armenia	123.35	121.16	118.36	115.56	113.21	111.47	110.21	109.34	108.67
Azerbaijan	87.37	88.71	90.07	91.29	92.35	93.29	94.19	95.23	95.23
Bahrain	738.63	736.26	754.5	773.45	793.99	816.7	842.02	870.5	902.68
Bangladesh	833.6	852.63	871.49	890.48	909.79	929.45	949.33	969.27	989.06
Bhutan	11.44	11.3	11.1	12.85	12.79	12.87	13.09	13.4	13.78
Brunei	50.16	51.61	53.08	54.54	55.98	57.39	58.78	60.13	61.44
Cambodia	52.82	54.72	56.69	58.66	60.58	62.44	64.23	65.95	64.23
China	122.58	124.09	125.52	126.95	128.34	129.69	131.02	132.29	133.44
Cyprus	84.76	86.65	88.62	90.6	92.57	94.51	96.4	98.27	100.16
Georgia	84.61	85.26	85.92	85.06	82.82	80.76	79.28	78.51	77.9
Hong Kong	5,810	5,859	5,961	6,096	5,863	6,129	6,180	6,232	6,292
India	298.84	304.88	310.94	317.03	323.18	329.37	335.6	341.86	348.1
Indonesia	101.91	103.65	105.36	107.06	108.72	110.36	111.98	113.58	115.17
Iran	35.18	35.69	36.11	36.11	37.04	37.64	38.33	39.06	39.78
Iraq	41.05	42.23	43.49	44.83	46.23	47.68	49.2	50.76	52.33
Japan	339.9	340.7	341.6	342.7	344	345	345.8	346.8	347.4
Jordan	42.5	45	47.5	49.8	51.8	53.4	54.8	55.8	56.8
Kazakhstan	6.09	6.09	6.05	5.96	5.86	5.77	5.68	5.58	5.53
Kuwait	114.2	108.25	102.3	96.35	90.4	91.6	96.3	103.1	109.8
Kyrgyzstan	23.3	23.5	23.5	23.5	23.8	24.1	24.5	24.9	25.2
Laos	19	19.5	20	20.5	21	21.5	21.9	22.3	22.7
Lebanon	269.1	275.8	283.6	290.8	296.5	300.2	302.3	304.4	308.6

[Table 03: Portion of Population Density Dataset]

3.1.4 Number of Refugee Countries

From the UNHCR Refugee Population data mentioned above, number of Refugee countries value was found out from year 1991 to 2016 by using the count function in Excel. 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 the independent variables.

3.1.5 Global Peace Index (GPI)

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) [11]. 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.

Country	score_2008	score_2009	score_2010	score_2011	score_2012
Afghanistan	3	3.358	3.252	3.212	3.252
Albania	1.91	1.89	1.925	1.912	1.927
Algeria	2.29	2.276	2.277	2.423	2.255
Argentina	1.77	1.846	1.962	1.852	1.763
Austria	1.291	1.24	1.29	1.337	1.328
Azerbaijan	2.29	2.342	2.367	2.379	2.36
Bahrain	1.8	1.815	1.956	2.398	2.247
Bangladesh	2.1	2.082	2.058	2.07	2.071
Belarus	2.06	2.046	2.204	2.283	2.208
Belgium	1.368	1.365	1.4	1.413	1.376
Bhutan	1.44	1.722	1.665	1.693	1.481
Bolivia	1.96	2.041	2.037	2.045	2.021
Bosnia	1.9	1.735	1.873	1.893	1.923

[Table 04: Portion of GPI Dataset]

3.1.6 Twitter Data

Initially keys generated from Twitter's developer account were used to write a python script to get tweets in English 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 means were used to get twitter historical data. TweetDeck was used instead as the other alternatives such as Gnip are not free of costs.

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 selected countries matching keyword 'refugee' for years 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 of and 6 columns of data. As the model was built for 20 countries, thus 20 datasets were prepared.

Year	GPI	GDP	No of country	Density	Population
1991		2845	2	56.9	13947
1992		3331	2	58.45	11299
1993		3665	13	59.96	154
1994		3971	17	61.08	5308
1995		4612	10	63.08	5278
1996		5103	10	64.71	5309
1997		4941	15	64.75	5285
1998		3470	18	66.38	50614
1999		3710	16	68.05	82501
2000		4287	16	69.7	50487
2001		4130	16	72.29	50466
2002		4380	14	74.27	50611
2003	1.566	4674	12	75.69	7424
2004	1.566	5171	19	77.1	24905
2005	1.566	5599	29	78.52	33693
2006	1.566	6264	29	79.94	37170
2007	1.566	7379	37	81.36	32658
2008	1.517	8647	35	82.78	36669
2009	1.52	7439	36	84.19	66137
2010	1.539	8920	37	85.59	81513
2011	1.485	10253	39	86.97	86667
2012	1.59	10655	43	88.33	90185
2013	1.574	10700	41	90.4	97513
2014	1.659	11009	44	92	99381
2015	1.561	9505	49	93.5	94166
2016	1.648	9374	48	94.4	92287

[Table 05: Dataset prepared for Model of Malaysia]

For sentiment analysis, tweets collected from TweetDeck were stored in a separate text file for each country. There were 2 columns, year of the tweet and the corresponding text. Table below demonstrates this arrangement of tweets.

Year	Tweets
2014	Hasina seeks amicable Refugee solution for Bangladesh
2014	BBS to run survey on undocumented Rohingya from Burma
2014	try to find the truth for Notorious BIG refugee
2014	Little hope for Burmese Rohingya in refugee camps
2014	Photograph of a Refugee within 2 miles of his homeland Refugees
2015	Dozens of dead near Yemen refugee camp
2015	refugee problems is the byproduct of western politics so they will affect badly.
2015	How does it feel to be a refugee?
2015	From Career Woman to #Refugee - The 1,500-mile journey of one Syrian mother
2015	Tusk calls for a U-turn in European refugee policy
2015	In Jordan, Zaatar's children: life in a refugee camp
2015	A refugee family's ordeal in Russia
2015	Did you know that Palestine refugees are the largest refugee population in the world?
2015	Life in a refugee camp: 'the cold and fear get in your bones'
2015	How other countries are responding to the UN's refugee requests
2015	US mayor under fire over refugee letter
2015	Europe anti-refugee rhetoric swells after Paris attacks
2015	Syrian refugee on Paris Attacks: "This is what we've been running from."
2015	Most Syrian refugee children miss formal education in Turkey
2015	Refugee children out of school constitutes about 40% of total out of school children. Sorry state of affair

[Table 06: Portion of Bangladesh's Twitter Dataset]

3.3 Imputation of Missing values

The GPI dataset had missing values as the ranking is done from 2008 to 2016 and the dataset for each country had row values for years 1991-2016. The imputation of missing values for each country's GPI column was done with mean**.**

Code:

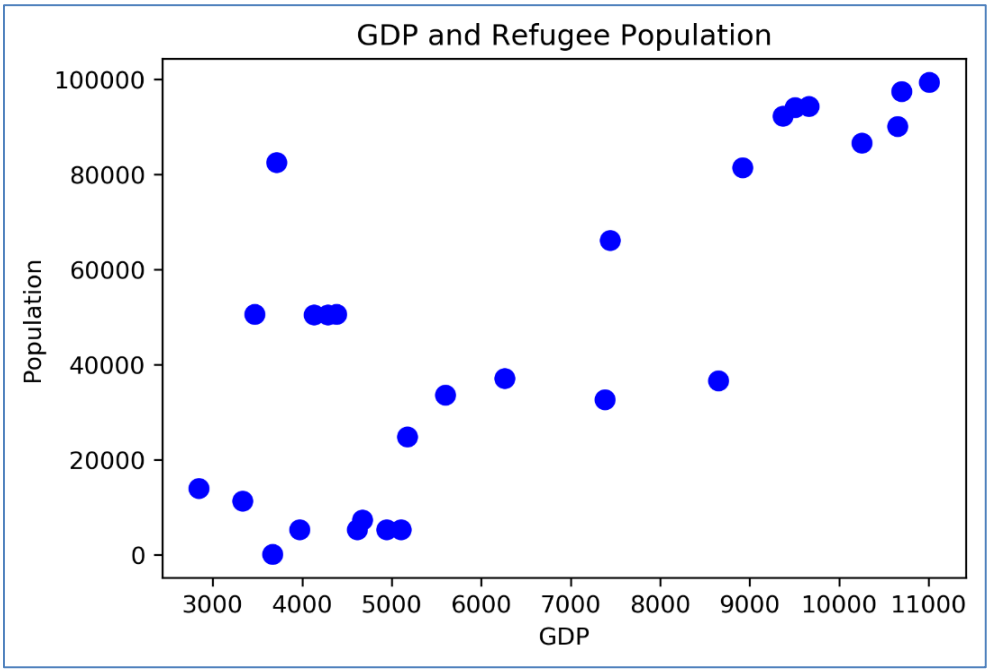
```
l = df['GPI'].mean()
df['GPI'].fillna(l,inplace = True)
```

3.4 Exploratory Data Analysis

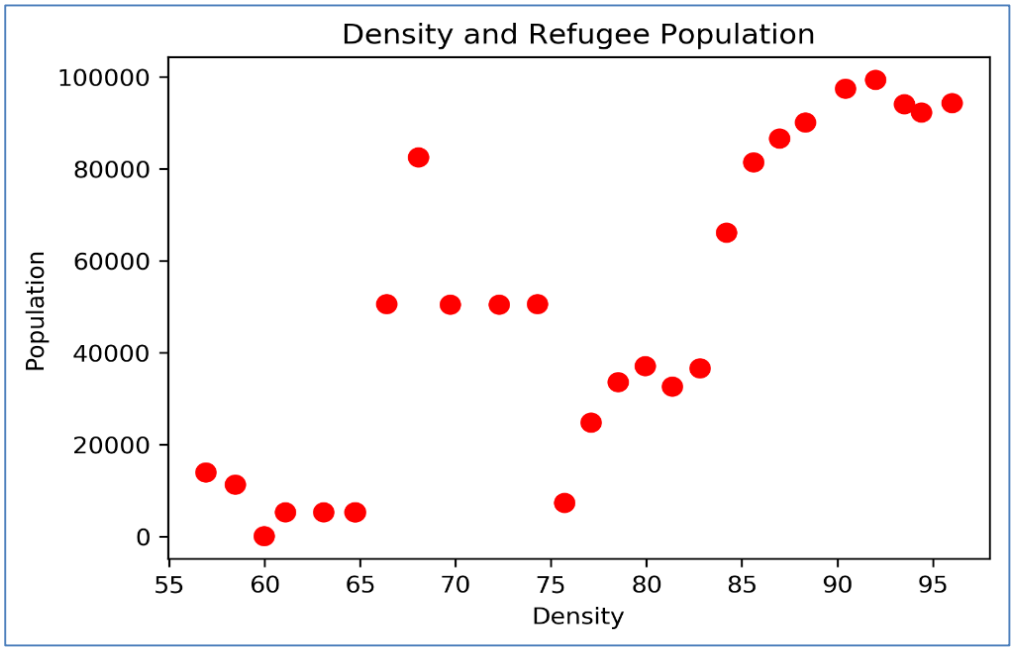
Initially the relationship between the dependent variable (Population of Refugees) and the independent variables (GDP, GPI, Population Density and No of Refugee Country) is analyzed through scatterplots. In the case of Malaysia, there appears to be a good linear relationship between the Population of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

Code:

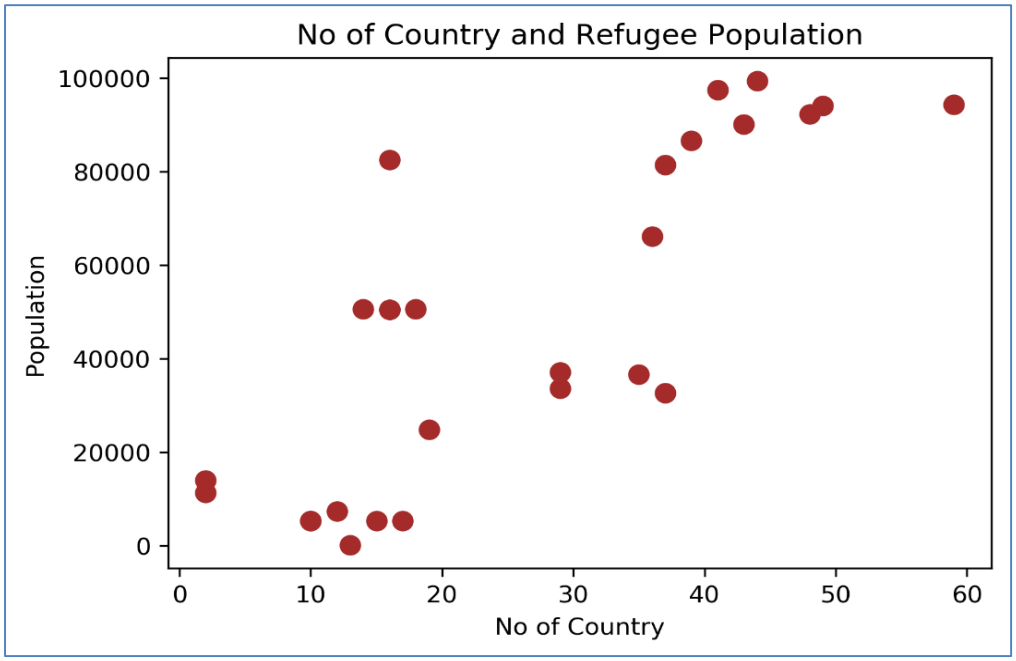
```
x = df['GDP']
y = df['Population']
plt.scatter(x,y,s=60,color='b')
plt.xlabel('GDP')
plt.ylabel('Population')
plt.title('GDP and Refugee Population')
plt.show()
```



[Figure 02: Scatterplot of GDP and Refugee Population of Malaysia]



[Figure 03: Scatterplot of Density and Refugee Population of Malaysia]



[Figure 04: Scatterplot of Number of Refugee Countries and Refugee Population of Malaysia]

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 (GDP, GPI, Population Density and No of Refugee Country). Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). The formula for linear regression is:

$$y = B0 + B1*x \qquad (1)$$

R squared value was used to check the linear relationship. R-squared is a statistical measure of how close the data are to the fitted regression line. The closer to

1.0, the better the fit of the regression line and the closer the line passes through all the points. Ideally the variables which gave values closest to 1.0 were selected for modelling.

Code for checking GDP and Population linear relationship:

```
feature_cols = ['GDP']
X = df[feature_cols]
y = df.Population
lm = LinearRegression()
lm.fit(X, y)

# print the coefficients
print("Intercept: ",lm.intercept_)
print("Coefficient: ",lm.coef_)

#check R squared value
print("R squared value: ",lm.score(X, y))
```

3.6 Modelling and Evaluation

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

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

The left-hand side of the equation (2) 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, Population Density and GPI 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 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. RMSE is the square root of the average of squared errors. The effect of each error on RMSE is proportional to the size of the squared error; thus, larger errors have a disproportionately large effect on RMSE. Consequently, RMSD is sensitive to outliers [12] [13]. MAE is simply the average absolute vertical or horizontal distance between each point in a scatter plot and the Y=X line. In other words, MAE is the average absolute difference between X and Y.

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.

Code for modelling:

```
# include key variables to build the model
X = df[['GDP', 'Density','No of country']]
y = df.Population

# Split data
X_train, X_test, y_train, y_test = train_test_split(X, y,test_size =
0.25, random_state=0)

# Instantiate model
lm2 = LinearRegression()

# Fit Model
lm2.fit(X_train, y_train)
```

```
# Predict
y_pred = lm2.predict(X_test)

# RMSE-Root Mean Squared Error
print("RMSE Score: ",np.sqrt(metrics.mean_squared_error(y_test,
y_pred)))

#MAE is the mean of the absolute value of the errors:
print("MAE Score: ",metrics.mean_absolute_error(y_test, y_pred))

#check R squared value
print("R squared value: ",lm2.score(X,y))
```

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, the 2017 to 2022 values were predicted 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.

Code:

```
#predict the GDP for future using a linear model with Year

X = df[['Year']]
y = df[['GDP']]

# Split data
X_train, X_test, y_train, y_test = train_test_split(X, y,test_size =
0.33, random_state=42)

# Instantiate model
lm2 = LinearRegression()
```

```
# Fit Model
lm2.fit(X_train, y_train)

#R squared value
print("R squared value: ",lm2.score(X,y))

# Predict
pred_GDP = [0,0,0,0,0,0]
for i in range (0,6):
    pred_GDP[i] = lm2.predict(2017+i)


#print predicted score
for i in range(0,6):
    print("Predicted GDP for year: ",(2017+i),pred_Den[i])
```

3.8 Prediction of Refugee Population

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

3.9 Sentiment Analysis

TextBlob library in python was used to analyze the sentiment of the tweets. Polarity score of the tweet was used to categorize the tweets into positive, negative and neutral sentiments.

The polarity score is a float within the range [-1.0, 1.0]. 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 yearly by dividing number of positive, negative and neutral tweets by total number of tweets respectively. This was done for all the years in our twitter dataset, from 2014 to 2017. The findings will be discussed in the next section.

Country	Number	Positive	Negative	Neutral
Bangladesh	105	41.9	14.28	43.8
India	111	27	17.11	55.86
Indonesia	64	37.5	18.75	43.75
Jordan	99	41.41	10.1	48.48
Pakistan	35	22.86	22.86	54.29

[Table 7: Portion of 2017 tweet findings. The tweets are arranged into positive, negative and neutral percentages using the steps mentioned above]

CHAPTER 4

Results

4.1 Model Results

4.1.1 Evaluation of Prediction

The models for each country predicted values up to the year 2022. As this is forecasting, 2022 refugee population values cannot be evaluated by comparison with real values. In this scenario the most reliable evaluation would be to check the values for 2016 prediction. The table demonstrates this for all the models built.

Country	Actual 2016	Predicted 2016
Malaysia	92,287	91,934
China	317,260	308,638
Yemen	269,796	243,703
Jordan	685,214	603,137
Thailand	106,471	114,070
Tajikistan	2,731	2,297
Japan	2,515	1,373
Indonesia	17,144	6,306
Turkey	2,869,420	1,688,008
Pakistan	1,352,572	1,567,471
Bangladesh	276,216	390,099
India	229,147	169,372
Iran	979,435	549,041
Nepal	25,250	59,519
Russia	229,017	31,952
Kuwait	945	3,770
Lebanon	1,012,984	703,453
Iraq	261,891	182,314
Armenia	17,897	-37,831
Kazakhstan	662	-292

[Table 8: Actual and Predicted 2016 values]

4.1.2 Prediction up to 2022

The table shows the values of prediction of each country's model from 2017 to 2022. Different countries experience different rates of increase and decrease according to the slope of their linear model.

According to the model findings, Turkey would have the highest population of Refugees in 2022 followed by Pakistan, Lebanon and Jordan.

Country	2017	2018	2019	2020	2021	2022
Malaysia	95,281	98,629	101,976	105,324	108,671	112,019
China	309,566	310,493	311,420	312,347	313,274	314,201
Yemen	253,711	263,719	273,727	283,735	293,744	303,752
Jordan	633,688	664,238	694,789	725,339	755,890	786,440
Thailand	114,425	114,780	115,135	115,490	115,845	116,200
Tajikistan	2,210	2,124	2,038	1,952	1,865	1,779
Japan	1,199	1,025	851	677	503	329
Indonesia	6,318	6,343	6,368	6,393	6,418	6,443
Turkey	1,781,358	1,874,707	1,968,057	2,061,407	2,154,757	224,8107
Pakistan	1,574,297	1,581,124	1,587,951	1,594,778	1,601,605	1,608,432
Bangladesh	407,191	424,284	441,377	458,469	475,562	492,654
India	167,000	164,628	162,256	159,884	157,513	155,141
Iran	476,236	403,432	330,627	257,822	185,018	112,213
Nepal	56,522	53,525	50,528	47,530	44,533	41,536
Russia	30,094	28,236	26,378	24,520	22,661	2,0803
Kuwait	3,260	2,750	2,240	1,731	1,221	712
Lebanon	744,210	784,967	825,724	866,481	907,238	947,995
Iraq	187,246	192,179	197,111	202,043	206,976	211,908
Armenia	-54,272	-70,714	-87,155	-10,3597	-120,038	-136,480
Kazakhstan	-1,092	-1,892	-2,692	-3,492	-4,292	-5,092

[Table 9: Prediction of Refugee Population in Asian Countries up-to year 2022]

4.2 Analysis of Prediction Results

The model predictions showed satisfactory results overall, especially with the available data. The 2016 predicted values for some of the models were close to the actual values but for few models it far from accurate.

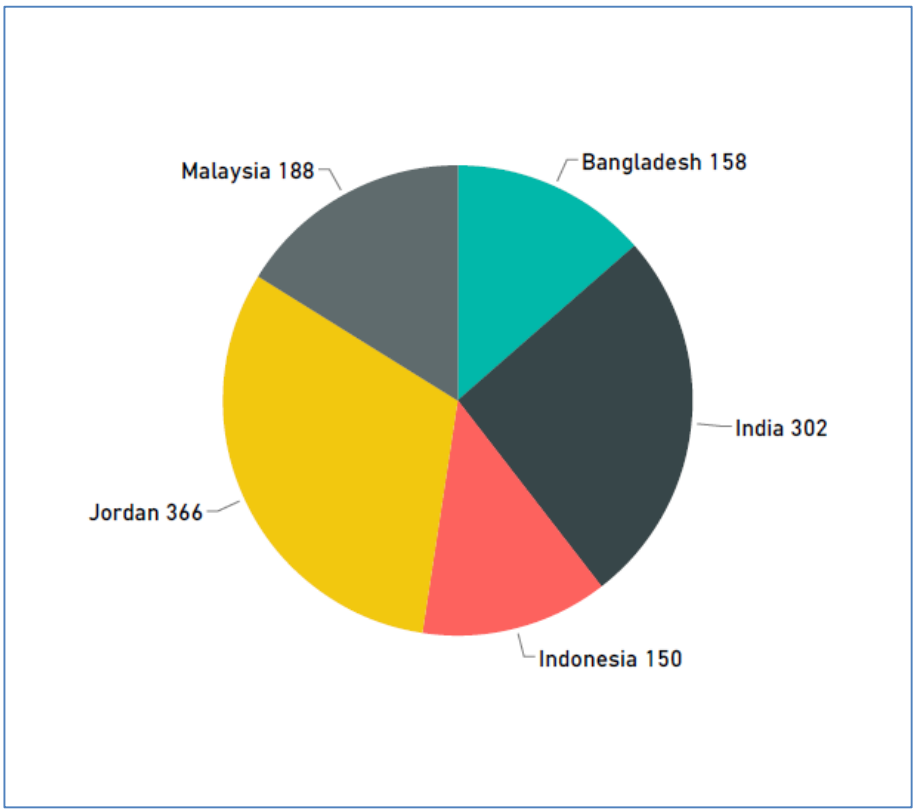
The instances where the relationship between the features (GDP, GPI, Number of Refugee Country and Density) and population of refugees followed a linear pattern, the results for 2016 predicted values came close to actual values. Forecasting of the independent variables impacted the model results. The countries for which the feature values increased/decreased linearly with year, the results were deemed to be more than satisfactory.

There were also 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. In some country datasets, there was a six-digit yearly refugee population value followed by a four-digit population value, with little change in the independent variables. Thus, in those cases, there was not a single variable which had R-squared value over 0.20, which greatly impacted the prediction accuracy. Feature selection was then carried out with the analogy of avoiding the worst variables. There were also instances where the forecasting of the independent values with year did not fetch good independent value predictions, as our assumption of forecasting was proved incorrect by ever-changing independent variable values. In these cases, the predicted refugee population did not show good results.

The suggested improvements would be discussed in the ‘Future Enhancement’ section in the next chapter.

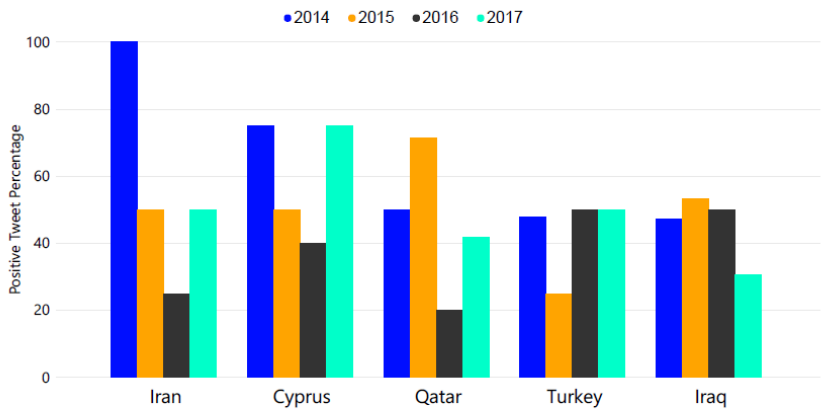
4.3 Sentiment Analysis Results

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

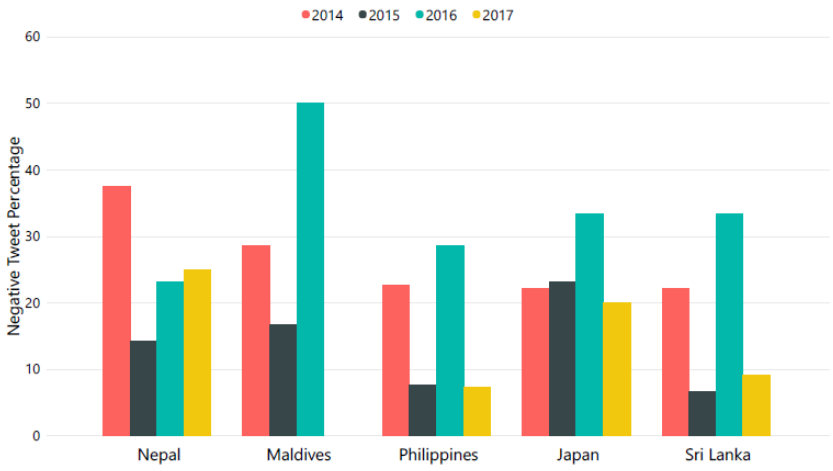


[Figure 5: Countries with most tweets about Refugees]

The yearly percentages for the top 5 countries with the highest percentage of positive and negative tweets are shown in figures 6 and 7 respectively.



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



[Figure 7: Countries with highest percentage of negative tweets]

CHAPTER 5

Conclusion

5.1 Correlation

After going through the findings, a good correlation can be drawn with the 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 both most number of tweets and top 5 highest predicted 2022 refugee population.

5.2 Future Enhancement

5.2.1 Improvement of Model

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 could be better predicted by adding the indicators as independent variables to the existing country datasets. The indicators of GPI mentioned above are currently not made open to the public.

Likewise, more relevant independent variables could be added for improvement of the model.

5.2.2 Alternate Prediction Technique

In this project, population of refugees entering countries were predicted. Enhancing this idea, population of refugees being produced from countries can be predicted. This would require prediction of violence and different sets of data needs to be used. This would have key significance, as steps can be taken to mitigate the cause of refugee production.

5.3 Conclusion

There were many limitations in carrying out this project due to lack of resources. Also, the allocated time for the FYP was only around 3 months which doesn't allow for any big scale project such as this. The results of the current models are very promising and can have massive impact on the refugee crisis. It is hoped that in the near future this is enhanced and used in the real world to benefit the governments, NGOs and refugees.

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

[3] 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.

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

[5] 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

[6] Brouckman, L, Wang, A., “Analyzing Twitter Sentiment on the Refugee Crisis in 2016”, Stanford, 2016

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

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

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

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

[12] Pontius, Robert; Thontteh, Olufunmilayo; Chen, Hao (2008). "Components of information for multiple resolution comparison between maps that share a real variable". Environmental Ecological Statistics. 15: 111–142.

[13] Willmott, Cort; Matsuura, Kenji (2006). "On the use of dimensioned measures of error to evaluate the performance of spatial interpolators". International Journal of Geographic Information Science. 20: 89–102.