UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity Dignified Life

Presidency of the Council of Ministers
Central Administration of Statistics

DATA-POP ALLIANCE
Changing the world with data

In collaboration with

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Webinar on Localized Big Data Applications for Supporting Decision-making

Lebanon Case: big data and official statistics - socioeconomic trends of refugees and host communities

Presented by: Dr. Ziad Abdallah

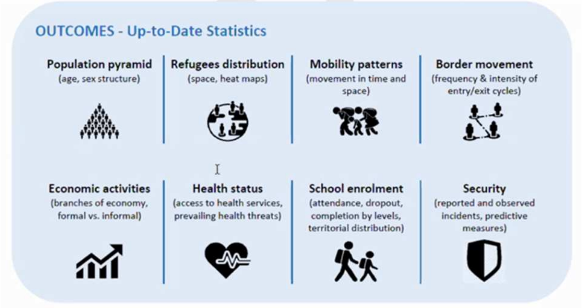
18 December 2020

KEY FINDINGS (not obvious in this field)

- Big Data Complements official statistics and cannot replace (confirmed)
- Big Data applications have Different Convention and usage (market): example Aid and Rescue in crisis response
- Proven usefulness for select indicators (CDR and FB population distribution)
- Less useful in certain indicators (FB and CDR Gender, Age Pyramid, and economic activities)
- Large data points needed to have regression from different sources (CDR and FB)
- Strongly recommended for crisis social media sentiments, and internet scraping (Twitter, GDELT)

Explored Human Development Outcomes

1. Population Pyramid
2. Refugee Distribution
3. Mobility Patterns
4. Border Movement
5. Economic Activities
6. Security
7. School Enrollment
8. Health Status



OUTCOMES - Up-to-Date Statistics

Population pyramid (age, sex structure)

Refugees distribution (space, heat maps)

Mobility patterns (movement in time and space)

Border movement (frequency & intensity of entry/exit cycles)

Economic activities (branches of economy, formal vs. informal)

Health status (access to health services, prevailing health threats)

School enrolment (attendance, dropout, completion by levels, territorial distribution)

Security (reported and observed incidents, predictive measures)

*not covered in this work due to nature of data

Methodology

Key steps:

1. Gathering data from traditional and non-traditional data sources
2. Creating simple indicators that can be used as covariates
3. Descriptive statistics
4. Simple models that show how covariates from non-traditional data sources can explain changes in traditional ones related to living conditions.
5. Testing out of sample predictions.

Data Sources

- Traditional - references
  - UNHCR Vulnerability Assessment of Syrian Refugees (VASYR) 2012-2019
  - CAS Labor Force and Household Conditions Survey (CASLFS) 2018
- “Non-traditional”
  - Call Detail Records
  - Facebook Advertising Platform (QCRI)
  - GDELT
  - Twitter

Call Detail Records (CDRs) Approach

Mobile CDRs: Calculated Outcomes

May-June-July of 2016-17-18-19; North and Bekaa only; 2 Operators

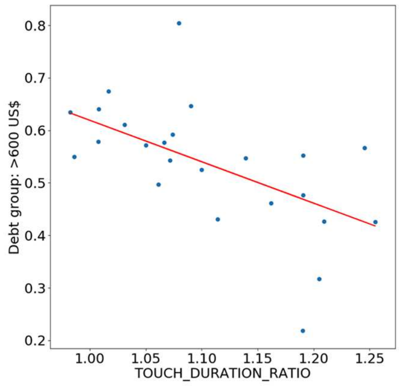
- Residents Distribution (total calls: Kaza / Muhafaza)
- Annual Mobility uses distribution (current - previous / previous)
- Population Pyramid (calls per age per gender)
- Border Movement (calls Tawasol trend across 3 years)
- Economic Activities:
  - Out & In Calls number, duration, ratio
  - Mobile Data Up & Down total and ratio
- Security: Density of Sites & Cells per Kaza

Population Distribution - Pearson Correlation

Target Variable	**Pearson Correlation**
Alfa Population Indicator	0.95
Touch Population Indicator	0.88
Touch Number of Cells	0.89
Alfa Number of Sites	0.81

Pearson correlation coefficient between CDR indicators and CASLFS population distribution

Economics - Regression with VASYR



Touch OUT duration to IN duration ratio vs. %families with debt >600$ Regression:

- R^2 Score: 0.36
- Mean Squared Error: 0.009

CDR Multivariate Regression

Target Variable	**K-Best Features**		**R^2 score**	**MSE**
	Feature	**Importance**		
CASLFS - Labor force participation rate (%)	**CDR** mobility	0.35	0.54	6
	CDR mobile data consumption download-to-total ratio	0.61		
	CDR outgoing-to-incoming call number ratio	0.03 (not significant)		
VASYR- % families with debt >600 US$	**CDR** outgoing-to-incoming call number ratio	0.22	0.35	0.009
	CDR outgoing-to-incoming call duration ratio	0.78		

CDR - Security

	Pearson correlation
Alfa population indicator	0.7
Touch population indicator	0.65

Pearson correlation between CDR indicators and VASYR security indicators

Facebook Approach

Data

- Estimates of Monthly Active Users (MAU) of Facebook.
- Syrian refugees in Lebanon were estimated (since FB not in Syria) as:
  - FB user over age 13
  - Expat (not born in Lebanon)
  - Arabic Speaking
  - Not from other FB registered Arab countries (e.g. Egypt, Morocco, Qatar, KSA, …etc)
- October 2019 and the attributes are:
  - **Age**
  - **Gender**
  - **Language**
  - **expat status**
  - **Education status**
  - **Device/Network types:** (3G/4G/WiFi, iOS/Android, latest Samsung/iPhones)

FB Demographics - Findings

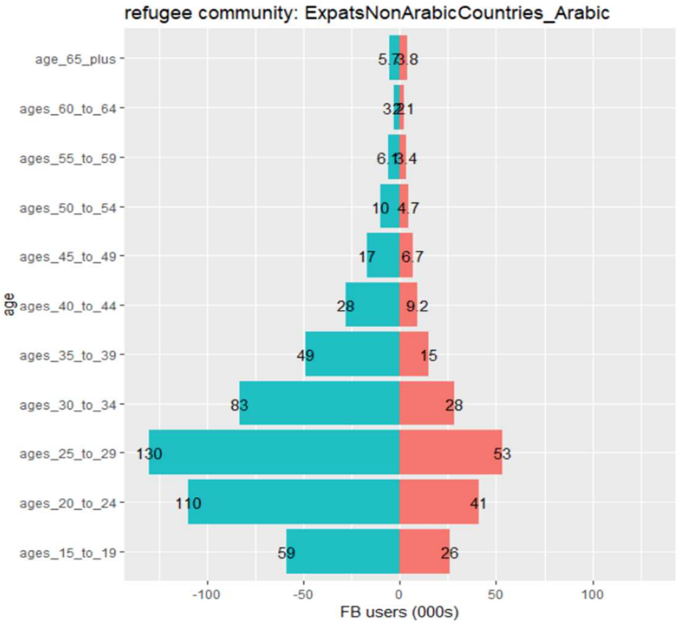
- Rounded Facebook monthly active users Oct 2019 (**age 13+)** compared to the numbers of registered refugees in UNHCR.

Governorate	FB users (host community)	FB users (refugee community)	Registered refugees (UNHCR report)	Ratio (FB users refugee community / Registered refugees)
Beqaa	330,000	130,000	342,875	0.379
North	490,000	100,000	243,125	0.411
Beirut	1,100,000	280,000	17,059	16.414 Outlier Capital City LARGE EXPAT
Mount Lebanon	290,000	110,000	210,950	0.521
South	300,000	52,000	66,929	0.777
Nabatieh	150,000	25,000	38,036	0.657

FB Refugees Gender Distribution

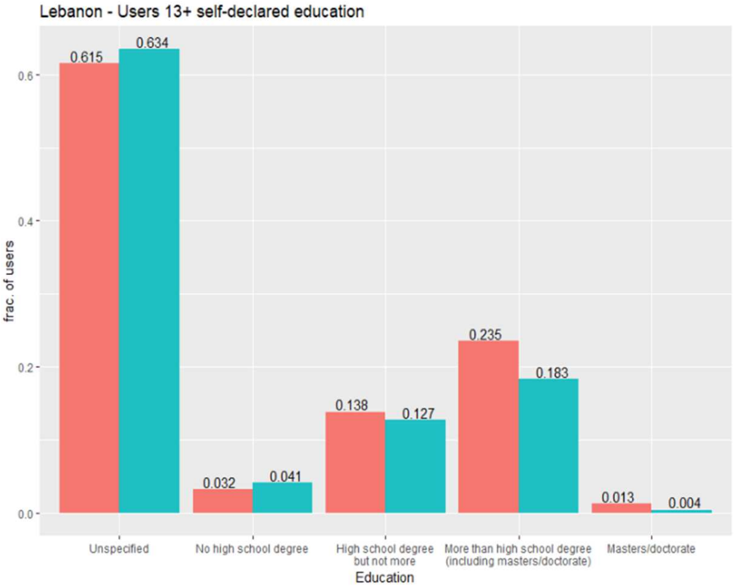
Graph showing refugee Community Age distribution by gender from Facebook

(number of male multiplied by – sign for visualization purposes)



FB Users: Educational Statuses

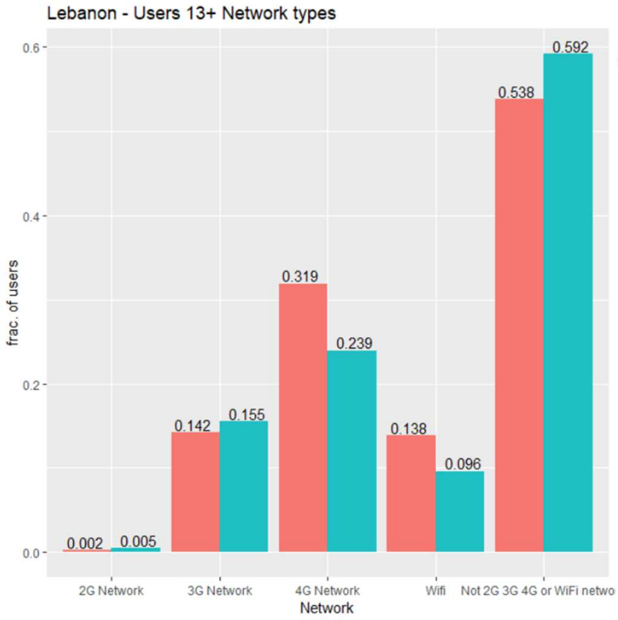
Graph showing self-declared education levels on Facebook Users in Lebanon for host (RED) and refugee (BLUE) groups



FB Users: Socio-economics

Access to devices/networks

Graph showing FB users in Lebanon by type of Network access for host (RED) and refugee (BLUE) communities.



GDELT

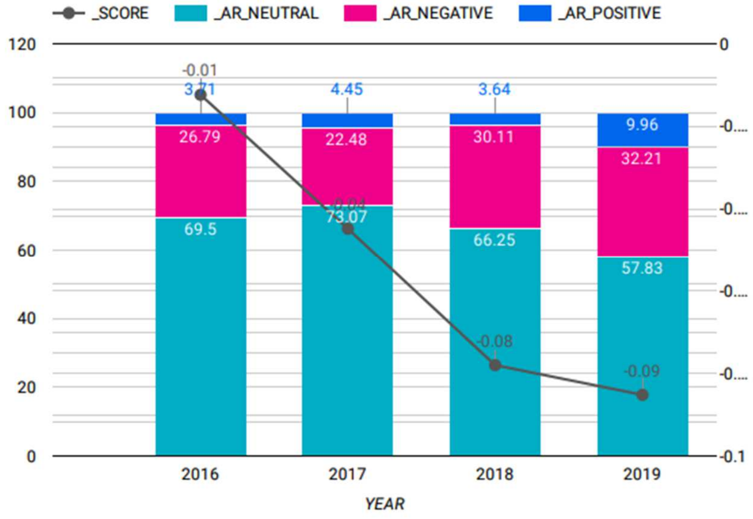
Media Sentiments: GDELT

- GDELT monitors the world's broadcast, print, and web news in over 100 languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events from each article.
- Extracted Arabic and English articles from GDELT reporting about events taking place in Lebanon involving Syrian refugees.
- GDELT provides a sentiment analysis score that ranges from -1 (negative) to +1 (positive).
- For Arabic sentiment analysis, open source API Mazajak was used that distributes percentage of the text as either positive, negative, or neutral.

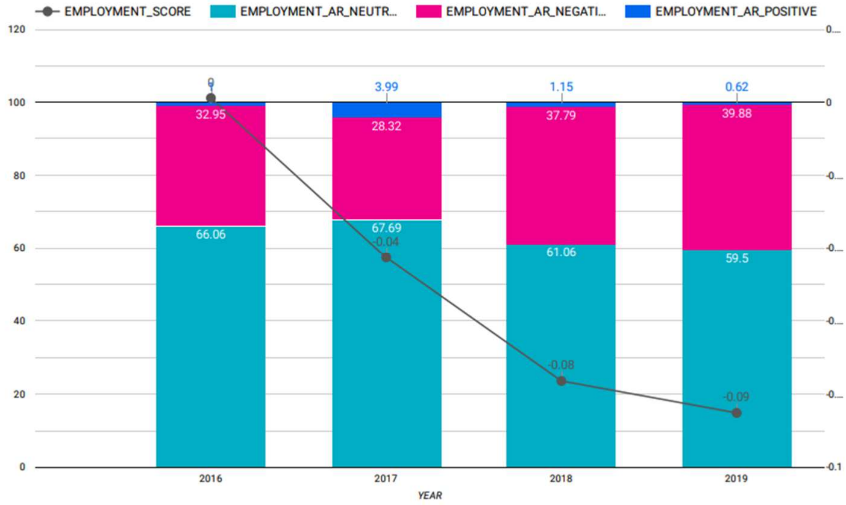
Articles About Syrian Refugees – Sentiment

Graph showing the distribution of the Arabic sentiment (% neutral, % negative, % positive) and English sentiment (ranges from -1 to +1) of the articles reporting about Syrian refugees from 2016 until 2019.

Total of 7400 unique articles



Articles About Employment - Sentiment



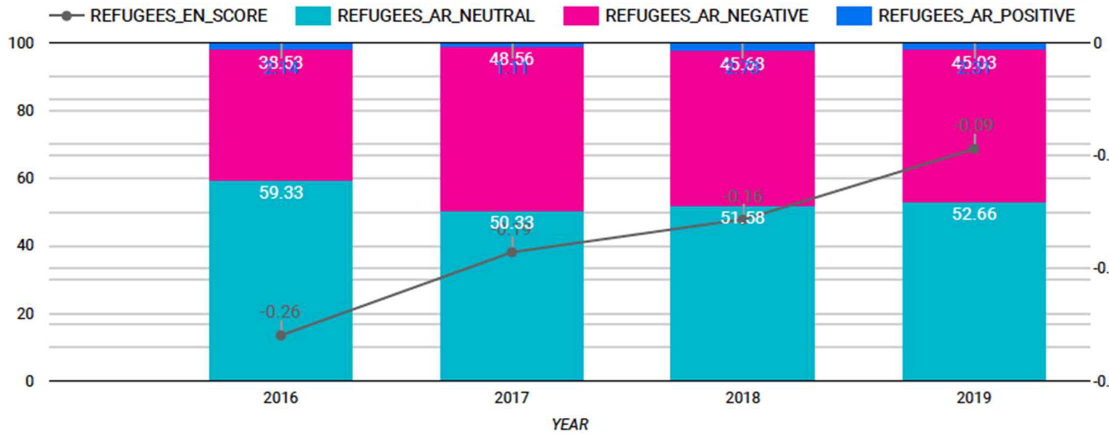
Graph showing the distribution of the Arabic sentiment (% neutral, % negative, % positive) and English sentiment (ranges from -1 to +1) of the articles reporting about Syrian refugees and mentioning employment from 2016 until 2019.
Total of 4466 unique articles

Twitter

Twitter Approach

- Similar to GDELT, extracted tweets from 2016 to 2019 using the same set of keywords and topics using Twitter Premium.
- Twitter premium allows 500 tweets per query.
- For each topic, two requests, one in Arabic and one in English, and dropped the duplicate tweets among the different keywords.
- The result was a sample of 4440 unique tweets about the Syrian refugees and the Syrian refugees in the topic of interest.
- Unlike GDELT, lack of geographical disaggregation on tweets.

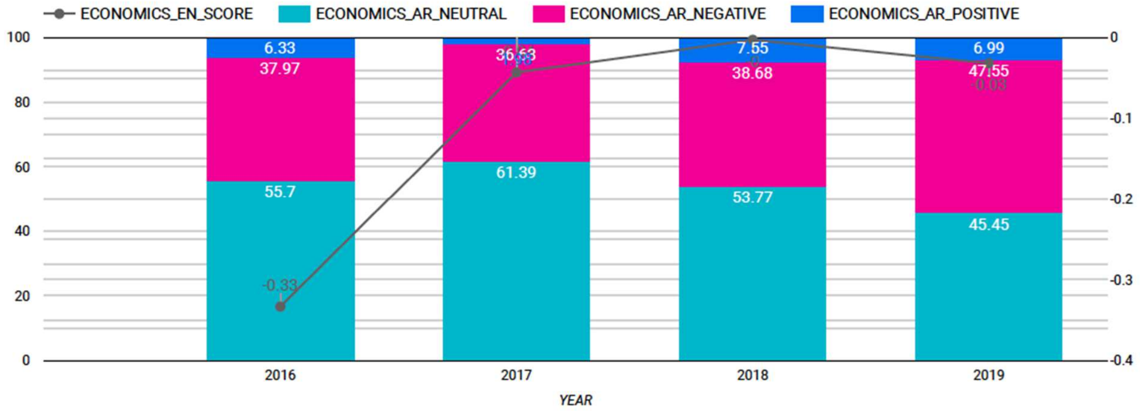
Tweets About Syrian Refugees - Sentiment



Graph showing the distribution of the Arabic sentiment (% neutral, % negative, % positive) and English sentiment (ranges from -1 to +1) of the tweets mentioning Syrian refugees from 2016 until 2019.
Sample Size:

- *Arabic tweets: 2126*
- *English tweets: 2314*

Tweets About Employment - Sentiment



Graph showing the distribution of the Arabic sentiment (% neutral, % negative, % positive) and English sentiment (ranges from -1 to +1) of the tweets mentioning Syrian refugees and economics from 2016 until 2019.

Sample Size:

- *Arabic tweets: 429*
- *English tweets: 405*

Discussions

Discussions

- In this Pilot project different non-traditional data sources were harnessed, particularly Mobile CDRs and Facebook Ad Platform in support of indicators relevant to 4 sustainable development goals (SDGs):

➢ Goal 1: End poverty in all its forms everywhere
➢ Goal 8: Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all
➢ Goal 10: Reduce inequality within and among countries
➢ Goal 16: Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels

Discussions

- A consistent gap between the actual number of Syrian refugees in Lebanon and those who have registered with UNHCR, hence only 44% of the overall eligible families for multipurpose cash assistance were provided with assistance

- Data gathered from non-traditional data sources would be key to improve the manner in which international and national humanitarian and development agencies identify populations and their movements, thereby enhancing their ability to assess assistance needs for Syrian refugees in Lebanon.

- The advantage of using available conventional ground truth data as a reference for the analyses is that weights can be given to the call behavior (calibration) so error compensation can be learned and properly corrected.

Discussions - Continued

- The correlations between the number of calls made and the population distribution were high, Facebook audience estimates correlated well with population proportions at the Muhafaza level, which can guide resource allocation.
- There was lower explanatory power in the economic models since having access to a phone introduces bias to the representativeness of call behavior for vulnerable populations, meaning that those who are more vulnerable are underrepresented.
- In spite of the high levels of data aggregation (same 3 Months for 4 years) and focus on two Muhafazas (Bekaa and North), the models were able to explain between 50 and 90 percent of the variation in available traditional ground truth labels (UNHCR and CAS) for Host Communities and Syrian Refugees especially in demographics distribution, and mobility.

Lessons Learnt

- Balanced between data richness (accuracy) and the protection of privacy.
- Significant political will, and strong partnerships developed in the framework of this project, made possible its completion. Special appreciation to the Ministry of Telecommunications for the confidence in CAS and ESCWA approach using Mobile Data records from local operators managed by the Ministry
- Furthermore, the CODE - Council for the Orientation for Development and Ethics - was a crucial body (Academic, Government, NGO, UNHCR, UNDP) to ensure the appropriateness and soundness of the research, as well as its ethical standards.
- Call Detail Records, safe research with strong privacy considerations, including on-site data processing, geographic and temporal data aggregation.
- Telecom Operators should be in the capacity of providing examples of their data structure, so that researchers can provide examples of code to create the indicators and aggregate them.
- Utilize readily made tools to create hundreds of different indicators that are often used in CDR research. Open source libraries that allow the creation of commonly used indicators in a relatively easy way.

Combining Different Data Sources

- Combining indicators gathered from the different data sources can further enhance the model's ability to predict the target values.
- However, a larger sample is required to make use of more indicators simultaneously.
- due to the aggregated nature of the data at hand, using a complex model combining numerous features increased the risk of overfitting.

Thank you!