“Social Media and Forced Displacement: Big Data Analytics & Machine-Learning”

White Paper
September 2017

UN GLOBAL PULSE | UNHCR INNOVATION SERVICE
# TABLE OF CONTENTS

- **ACKNOWLEDGMENTS** 4
- **SUMMARY**
- **BACKGROUND** 5
- **PROJECT OVERVIEW**
  - A COMPENDIUM OF MINI-STUDIES USING SOCIAL MEDIA DATA 6
    - Queries and Taxonomies
    - Classification
    - Iteration 1
      - Table 1: Initial Monitors Overview
    - Hypotheses
    - Setup
    - Insights
    - Iteration 2
      - Table 2: Situational Awareness Monitors Overview
    - Setup
    - Categories
    - Insights 10
- **LIMITATIONS AND LESSONS LEARNED** 11
- **THE WAY FORWARD**
- **ANNEXES** 13
  - Annex I: Data Query Taxonomies per Hypothesis
  - Annex II: Tweets found and catalogued by AI 16
  - Annex III: Data Visualizations (Quantitative inputs) 19
  - Annex III: Data Visualizations (Qualitative inputs) 21
  - Annex IV: Interactive map 23
ACKNOWLEDGMENTS
This paper was developed by colleagues from UNHCR Innovation Service and UN Global Pulse, an innovation initiative of the United Nations. UN Global Pulse would also like to thank the Government of the Netherlands for supporting its network of Pulse Labs and the activities under this project.

SUMMARY
This white paper summarizes the initial findings and lessons learned from a project conducted by UNHCR’s Innovation Service and UN Global Pulse to inform on the viability and value of social media analytics to complement understandings of the Europe Refugee Emergency.

Ongoing conflicts and violence around the world led over 1.4 million people to seek refuge in Europe between 2015 and the first part of 2017. Data from social media offers a wealth of information that can be parsed to better understand what people think, and how people feel about things affecting their lives, such as the displacement and movement of large volumes of people. Researchers in turn, can use this data to inform topics of interest; decision makers can use such data as evidence on which to inform for example, programmatic responses and alterations.

The paper outlines the process, questions and methodology used to develop the project and presents preliminary observations on how aspects of the Europe Refugee Emergency are related on Twitter. The paper describes ten quantitative social media mini-studies that were developed as part of the project.

The project team initially set out to explore the value of social media both for monitoring Persons of Concern’s (PoC), sentiment towards the provision of services, and their interactions with service providers. However, based on inconclusive initial results and anticipating an increase in negative public views towards PoC following the 2015-2016 terrorist attacks in Europe, the project refocused on the analysis of host communities’ sentiment towards PoC in reaction to incidents taking place in different European countries. Findings revealed that within local active Twitter communities, a small number of people connected PoC and the different terrorist attacks. Being able to assess peoples’ views in real-time provides a unique opportunity for UNHCR to counter non-conducive behaviour online. It also allows the Agency to better understand generalized perceptions vis-a-vis longer-term solutions for PoC.

The processes detailed herein are intended to serve as examples and to inspire other agencies looking to use social media and data analytics to inform decision-making processes, operational responses and policy development in emergency-related contexts.

Keywords: social media monitoring, big data, big data analytics, machine-learning, artificial intelligence, data science, refugees, asylum-seekers, migrants, Europe, sentiment analysis, xenophobia, data science.

BACKGROUND
The Europe Refugee Emergency was a constantly, and rapidly-changing context. Ongoing conflicts and violence around the world led over 14 million people to seek refuge in Europe between 2015 and the first part of 2017. This included increasing numbers of families, women, and unaccompanied and separated children—some seeking to reunite with other family members already in Europe. This new movement was challenging for many organizations, including UNHCR; people moved quickly, often across several international boundaries in very short periods of time; sometimes encountering changing protection risks, particularly when legal practices evolved, when borders closed, or when alternative routes begin to develop.

According to a report released by Social Media for Good, social media monitoring can provide significant value for decision makers in such dynamic contexts, where humanitarian access is poor, the information landscape fragmented, and social media widely used. For example, UNHCR’s report “From a Refugee Perspective” portrays the discourse of refugees and migrants and the use of social media. Social media platforms are powerful communications tools for humanitarian organizations, both at a strategic corporate level, and as a mechanism for communicating with affected communities. They also contain a wealth of information that can be parsed to measure and monitor conversations and emerging narratives.

Further, sentiment analysis of social media content can be used to capture public perceptions of an organization and its activities in a particular context to not only help develop new strategies, but also to ensure that existing programmes and projects are re-aligned and course-corrected in real-time. Several pilot and research projects have shown the feasibility of using social media data to crowdsourc topics of relevance to sustainable development and humanitarian action. However, there has been little effort in extending the quantification of online sentiment to inform on interactions between PoC and services providers. Similarly, organizations would benefit from understanding how host communities view PoC on social media to inform their decision-making processes.

UNHCR currently uses social media for two main purposes: 1) to publicly portray the Agency’s work, and digitally engage with public audiences; and 2) to communicate with affected populations. While UNHCR has a strong presence among different audiences on platforms, such as Twitter, Facebook, and Instagram, and it has clear guidelines on the use of these platforms for communication purposes. While the use of social media for CwC-supported activities is relatively new to the Agency, there are many promising efforts underway, both at Headquarters, and in field operations.

PROJECT OVERVIEW
The work described in this paper was initiated by the UNHCR Winter Operations Cell and UNHCR’s Innovation Service in November 2015. UNHCR recognized that big data analytics could provide additional insights into understanding the protection environment within the Europe Refugee Emergency. However, it did not have vast in-house knowledge, skills, or the necessary tools to conduct large-scale analyses. Therefore, it was limited in its ability to feed potentially valuable information contained in big data into operational responses.

To validate the value of social media data in emergent situations, UNHCR’s Innovation Service partnered with UN Global Pulse in January 2016. UN Global Pulse provided technical guidance, coaching and tools for the project. The joint-collaboration explored how alternative sources of data can and should play a role in pursuing humanitarian outcomes.

The project team identified two opportunities in which social media could be harnessed to better understand the Europe Refugee Emergency. Q₁: Monitor interactions between PoC, and between service providers and PoC, in an aggregated form, and; Q₂: Understand the sentiment of PoC, host communities to the extent which PoC have transited, in aggregated form.

The project envisioned a near-real-time monitoring system that could inform operational responses in support of the Europe Emergency Regional protection strategy. This system would have a two-tier architecture, with a machine learning component in the
The ten studies were divided into two main iterations. For each of the iterations, a methodology largely inspired by the Harvard Data Science curriculum was used. The following sections detail the data and tools that were employed, discuss the main hypotheses, and share the general iterative procedure.

## Data and Tools

**Twitter posts**, or tweets, are mostly public expressions of ideas and opinions\(^6\) as opposed to Facebook posts, which are mostly private. As of 2014, only 5% of Twitter accounts are protected\(^7\). Therefore, given the majority of Facebook posts are private, and potentially a PoC could be the one expressing an opinion, the project chose tweets as the main source of data, complying with UNHCR’s data protection policy\(^8\).

UN Global Pulse has a long-term research partnership with Crimson Hexagon that allowed the project team to use the company’s ForSight tool to access and analyze social media posts. Crimson Hexagon provides an online social media monitoring platform that enables users to create monitors, which have built-in machine learning capabilities to semi-automatically classify and extract sentiment from posts. These capabilities are based on algorithms that are iteratively improved using a training dataset, i.e., a curated collection of posts that help train the monitor to correctly interpret any new incoming posts\(^9\).

The three main steps in setting up and training a monitor are: 1) defining a taxonomy to identify the keywords, hashtags, and phrases that will help retrieve the most relevant posts from the social media platform of interest (e.g., Twitter), 2) formulating a query using those terms to retrieve the posts, and 3) manually classifying an initial subset of the retrieved posts to establish the training dataset. Once a monitor is trained, it provides different views of automatically classified posts, as well as the sentiments extracted from them, which enable users to conduct a variety of quantitative analyses. While the project described in this paper was implemented with the ForSight tool, the methodology is generic and can be executed with other technological solutions.

**Queries and Taxonomies**

Formulating appropriate queries is not always straightforward, and a significant amount of effort can be put into training a monitor before an inadequate query is detected. This is an iterative process that involves a certain degree of trial and error. For example, selecting the appropriate vocabulary can be difficult. Tweets abound with colloquial language and “internet-speak”—Arabic slang typically varies across countries and regions, and can be written in either Arabic alphabet (which is diakritized\(^10\)), or using the Roman alphabet (e.g., ArabEasy, which is sinistradextran\(^11\)). The 140-character restriction on Twitter also encourages word abbreviation. As a rule of thumb, a lack of relevancy can be indicative of a poor query, or a very restrictive combination of keywords based on the use of logical operators (AND/OR/NOT).

In addition, certain assumptions must be made regarding tweeters’ specific knowledge of the topic of interest. For example, a premise of this project is that, in general, people who tweet have little to no knowledge of the legal and protection differences between migrants and refugees. Both terms were used as synonyms in the queries, even though they have different implications for UNHCR. Contrary to migrants, refugees are specifically defined and protected by international law, particularly regarding refoulement\(^12\).

Finally, queries can be restrained in space and time. These two dimensions can be helpful for further bringing out the voice and opinions of, for example, PoC vs. host communities. Geo-referencing of social media posts\(^13\) can be done based on a combination of the location declared by the user in his/her profile, and the latest location(s) from where s/he posted.

### Classification

The classification process first requires determining a set of relevant categories, in which the queried posts will be filed. In initial explorations, the project team found that simple dichotomous categories are more effective, like racist–non-racist, or positive–negative.

| Categories for irrelevant and neutral posts are also useful, since all posts may not fit into the dichotomous pair, either because the context is inapplicable, or because it is incongruence\(^14\). All the categories determined in the mini-studies are presented in Annex I, along with their respective queries.

### An initial subset of posts must then be filed manually to create the training dataset (or training tweets), which the monitor’s underlying algorithms will use to automatically classify new incoming posts. This involves personal judgment as to whether content is relevant or not, and can turn out to be a lengthy procedure. Typically, the more categories there are, the more posts there are that need to be read, and manually sorted.

**Iteration 1**

The project team conducted six mini-studies in this iteration. For each study, a unique monitor on Crimson Hexagon was created (see Table 1).

<table>
<thead>
<tr>
<th>Table 1: Initial Monitors Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monitor</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>1. Interactions Arabic</td>
</tr>
<tr>
<td>2. Interactions Farsi</td>
</tr>
<tr>
<td>3. Xenophobia Greek</td>
</tr>
<tr>
<td>4. Xenophobia English</td>
</tr>
<tr>
<td>5. Xenophobia Arabic</td>
</tr>
<tr>
<td>6. Xenophobia Farsi</td>
</tr>
</tbody>
</table>

### Source:

Crimson Hexagon ForSight tool

### Hypotheses

**Hypothesis for O1**

The assumption for O1 was that analyzing social media posts could provide insights into, for example, altered routes, or the conversations PoC are having with service providers, including smugglers; and that this could provide better situational awareness for decision-making, and thereby better inform the orientation of resource allocations, and advocacy efforts.

**Hypothesis for O2**

Understanding PoC and host communities’ mutual sentiments will reveal how both groups view and react to asylum conditions and protection. This can inform programme design and planning strategies for Strategy Objective 3 of the Europe Emergency Regional Protection Strategy: Access to territory and asylum is safe.

The assumption for O2 was that understanding the sentiment that host communities express on social media could help identify pockets of, for example, xenophobic attitudes towards PoC, and that this could help UNHCR to improve the conditions in which durable solutions may occur, by better targeting corporate communications, and advocacy-related actions around legislation in specific countries.
The flexibility of language use on social media requires native speakers to c ommunicate and classify the train- ing tweets. Native speakers alone can understand the semantic nuances, colloquial, and even unusual uses of local language, as well as typical abbreviations discussed above. The project team relied on a group of English, Farsi, Arabic, and Greek native speakers, all of whom have basic knowledge in computer pro- gramming, as their paper will refer to the team of native speakers as the monitors. The monitor trainers were coordinated by UNHCR’s Innovation Service’s Data Scientist, and overseen by UN Global Pulse’s technical team.

The project team concentrated on Greece for the first iteration, as the country has played many different roles throughout the Europe Refugee Emergency—it is now a host country for a static population of refu- gees and migrants. Based on UNHCR data from the Europe Regional emergency—which includes demo- graphic population data—it was assumed that PoC would largely be Arabic or Farsi speakers, and that posts from Greece in these languages would likely be those of PoC.

To address O₁H, the trainers set up two specific monitors to track interactions between PoC, service pro- viders, and the general public regarding access to territory, asylum conditions, shelter conditions, trans- portation, and movement in Greece. The first monitor was called, as expected, Interactions Farsi. The second monitor was set up for Farsi, and was called Interactions Arabic. A full description of the categories and queries used is provided in Annex I. Tweets were retrieved from February 1st, 2015 or February 1st 2016 to April 18th, 2017 (end of the study). February 2015 corresponds to a period of major influx of PoC in Europe. February 1st, 2016 is approximately one month prior the EU-Turkey Agreement.

To address O₂H, the trainers set up four monitors to track negative sentiment and perceptions, like xenophobic, discriminatory, or racist sentiments, of host communities towards PoC in Greek, English, Arabic, and Farsi. The monitors were respective- ly called Xenophobia Greek, Xenophobia English, Xenophobia Arabic, and Xenophobia Farsi (see Annex I). Tweets were monitored for the period February 1st, 2015 or February 1st, 2016 to April 18th, 2017.

To build the Xenophobia monitors (O₂H), the project used the following categories to classify the tweets:

Xenophobic: tweets that express negative attitude, perception, or belief about PoC;
Neutral: tweets that describe facts about PoC (for example, news articles) but that do not express a strong sentiment or any sentiment at all;
Irrelevant: tweets that are not related to PoC.

The monitor trainers identified posts as belonging to the xenophobic category based on the UNESCO’s definition of xenophobia: “behavior of hostility based on existing racial, ethnic, religious, cultural, or national prejudice”; and the UN Fund for D161994 Foreign Concepts of Slavery. OHCHR, declared the defini- tion of xenophobia: “attitudes, prejudices, and behavior that reject, exclude and often vilify persons, based on the perception that they are outsiders or foreign to the community, society or national identity”.

The monitors also distinguished between factual, opin- ion-driven, rumor-driven, and breaking news tweets, in order to adequately train the machine for the neutral category. They further subtracted re-tweets (RT) from certain queries, following the findings of Mendoza et al., especially for the monitors related to O₁H, to avoid “inflating” the number of xenophobic posts.

The Interactions Arabic monitor was successfully trained but did not retrieve a large number of rele- vant tweets (<7,000—see Table 1). Annex II portrays a subset of these posts. The Interactions Farsi monitor, however, could not be trained, due to an apparent lack of tweets in Farsi (<1,500) regarding access to territory, shelter conditions, and transportation. These results did not provide enough data to confirm or refute O₁H, and could indicate that PoC—assumed to be either Arabic or Farsi speakers—neither do not use Twitter to inquire about, complain, or request ser- vices; b) do not have access to Twitter; or c) prefer other communications channels. The latter two pos- sibilities seem further supported by the Xenophobia Arabic, and Xenophobia Farsi monitors, which also retrieved a very low number of tweets (<200 and <150, respectively).

The analysis also showed it is difficult to systemati- cally separate tweets coming from PoC, host com- munities, and the general public for further analy- sis. Only few tweets described access to territory in Europe—including closing borders and entry restric- tions—anyway conditions, and the economic chal- lenges encountered during, and at the end of their journey, while many expressed the sentiment of host communities towards PoC. In hindsight, this could have been caused by improper querying and training of the monitors. Based on these early insights, the project decided to concentrate on O₂H.

The analysis of the O₁H monitors found few online signals for the Arabic and Farsi monitors. For English and Greek however, the number of posts was much bigger, in the order of thousands. Interestingly, only 5% of the tweets retrieved by the Xenophobia Greek monitor (12,423 out of 248,691—see Table 1) were clas- sified as xenophobic, compared to 15% (3,969 out of 26,466) in the Xenophobia English monitor. Although the monitors (queries) retrieved a larger number of posts in Greek, the analysis did not reveal the abso- lute number of tweets. However, with the sample re- trieved, there were more xenophobic posts in English than in Greek for this particular geographic location. See Annex 3 for a summary of the main topics dis- cussed in tweets retrieved by the Xenophobia Greek monitor.

Iteration 2

Four follow up studies were conducted in the second iteration. The project created a unique monitor for each study using Crimson Hexagon (see Table 2). The project team focused on measuring the volume of posts that either blamed, or defended PoC to gauge public opinion and understand whether opin- ions were generally in favour or against PoC.

The hypothesis for O₂H: Host communities and the general public may make a link between PoC and terrorist attacks.

Setup

To address O₂H, the monitors created four additional monitors that covered the unforeseen inci- dents in Nice (FR), Munich (DE), Saint-Étienne (FR), and Berlin (DE), which occurred on the 14th, 22nd, and 27th of July, and on the 18th of December 2016, respectively.

Each was intended to gauge responses to the ter- rorist attacks, and how these might be related to PoC in the global Twittersphere. They were respec- tively called Situation Awareness Nice, Situation Awareness Munich, Situation Awareness Saint-Étienne, and Situation Awareness Berlin. All were trained in English, French, Greek, and German—except the Situation Awareness Saint-Étienne, which was only trained in English and French—using almost exactly the same query—only some local ref- erences, and particular hashtags specific to each incident varied. Particular attention was given to employing the same vocabulary for each language to enable a relative degree of comparison between monitors. For example, “attack” in English was trans- lated to “attaque” in French, “επίθεση” in Greek, and “Anschlag” in German.

The monitors were not restrained to specific geo- graphic boundaries, but rather looked to under- stand global reactions and opinion. Nevertheless, the choice of language did concentrate the tweets

<table>
<thead>
<tr>
<th>Monitor</th>
<th>Unit of Analysis: Geography</th>
<th>Unit of Analysis: Timeframe</th>
<th>Language</th>
<th>Number of Posts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Situation Awareness Nice</td>
<td>Worldwide</td>
<td>Date of event (14 July 2016)</td>
<td>English, French, German</td>
<td>3,748,198</td>
</tr>
<tr>
<td>Situation Awareness Munich</td>
<td>Worldwide</td>
<td>Date of event (22 July 2016)</td>
<td>English, French, German</td>
<td>58,815,918</td>
</tr>
<tr>
<td>Situation Awareness Saint-Étienne</td>
<td>Worldwide</td>
<td>Date of event (27 July 2016)</td>
<td>English</td>
<td>28,884,522</td>
</tr>
<tr>
<td>Situation Awareness Berlin</td>
<td>Worldwide</td>
<td>Date of event (08 December 2016)</td>
<td>English, French, German</td>
<td>353,580,956</td>
</tr>
</tbody>
</table>

Source: Crimson Hexagon Foresight tool

Hypotheses

Based on the first iteration, and in reaction to incon- clusive results as well as in reaction to a number of terrorist attacks which occurred in Europe—resulting in refugees being mentioned in various media, includ- ing social media, in potentially concerning ways—, the project refocused to explore whether social media could provide a way to:

Q2: Monitor the general public’s opinion on possible mislead relations between PoC and terrorist attacks, in aggregated form.

Table 2: Situational Awareness Monitors Overview

1 Number of posts analyzed by the machine to the date: April 18th, 2017.
retrieved to areas where those languages are spoken (see Annex I for details on the categories and queries that were used for each language). Tweets in each monitor were filtered onwards from the date of the incident covered, i.e., in the aftermath of the terrorist attack, until April 18th, 2017.

Categories

The following categories were used for the situation awareness monitors (O:\H)

Blame: tweets that explicitly blame PoC for the incident.

Don’t Blame: tweets that advocate for not blaming PoC for the incident, or at least that attempt to de-attach them.

No reference to PoC: tweets that describe facts about the incident, but that do not mention PoC.

Irrelevant: tweets that mention PoC, but that are not related to the incident.

Off-topic: tweets that are neither related to PoC, or the incident.

Insights

While the Situation Awareness Saint-Étienne monitor gathered a significant number of tweets (>28M—see Table 1 for details of the incident tended to skew the results: PoC were rather linked with fundamentalism, than with the event itself. As such, this monitor was discarded, and the project team further investigated the incidents that did not specifically target religion.

Only 6% of the tweets retrieved by the Situation Awareness Nice monitor, 11% by the Situation Awareness Munich monitor, and 5% by the Situation Awareness Berlin blamed PoC for the incident. There were also more blame tweets in the Situation Awareness Berlin monitor than in the other two, with 7% of posts expressing explicit support for PoC in German, condemning racism and xenophobia, and stating that terrorism and violence are the main reasons why PoC flee their homes in the first place. It is also important to note that while the percentages of posts blaming refugees for the incidents are small, they still represent between hundreds of thousands to several millions of spontaneous messages of this direction: 0.2M tweets (Nice), 6.4 M tweets (Munich), and 17.6 M (Berlin).

The Situation Awareness Berlin monitor retrieved a significantly higher absolute number of tweets connecting PoC with the attack. These results could be attributed to several instances. First, because the police quickly identified and arrested a Pakistani asylum seeker as the perpetrator of the attack.

Although he was later released when found innocent, Twitter users pursued the discussion on a possible relationship between the incident and PoC. On December 20th, the police arrested the Pakistani suspect, even after the police had clarified that there was a Tunisian suspect. Second, this was the third in a series of recent incidents in Germany, after the Munich attack, and the Berlin monitor, where there were many sarcastic references to the radio interview with Angela Merkel.

The project initially imagined it would be possible to filter out the voices of PoC by collecting posts in Arabic and Farsi, and by geo-locating their point of origin. However, they turned out to be extremely im-possible to detect because the majority was indeed a migrant or refugee, or simply a person from the host community or local diaspora—especially since the Arabic and Farsi monitors in the first iteration did not retrieve many tweets. This would induce a high degree of uncer-tainty in any attempt to address hypotheses related to PoC vs. host communities, which is why a deeper analysis of PoC was not employed. However, extend-ing this research to other social media platforms like Facebook, which PoC use extensively 21, might help facilitate the distinction, and help better understand interactions.

The second iteration also showed the extent to which the comprehensiveness of the vocabulary used in a query could both increase the volume of retrieved posts, and their relevancy. By and large, several iterations are needed to capture all the information. Secondly, the messages should be manual-ly scanned to be able to re-classify and re-train the machine to include relevant words. Although it was generally similar, the query used for the Situation Awareness Berlin monitor was more sophisticated, and better tailored, than those used for the other situ-ation awareness monitors (see Annex I). This was the result of immediate feedback received from end-us-ers within Germany on specific language nuances. The posts were thus re-classified, which was also used by local media and the general public when referring to PoC. UNHCR’s Innovation Service’s Community and Content Manager also assisted, pointing to specific hashtags and keywords that were being used.

More generally, findings showed that working with social media requires a dynamic mindset. The project had to adapt and iterate rapidly. The hypotheses from the first iteration proved too broad, and for them to be of any tangible use to UNHCR, they had to be adjust-able. They needed more resources than the project ini-tially identified within UNHCR, as linking social media monitoring with operational responses and planning is a new concept for the Agency—the mini-studies were typically inaccurately conflated with a range of other social-media-driven projects, including CwC efforts, Information Management work, Communication and Social inclusion (PI) activities, and UNHCR digital brand marketing.

In addition, while this white paper refers to the current work as “mini-studies,” UNHCR has em-phasized this was a labor-intensive process. UNHCR’s Innovation Service had to resolve to recruit monitor trainers, all of whom had both relevant native language skills, and basic knowledge in computer program-ming. UN Global Pulse also invested the equivalent of a full-time staff member, in addition to providing the necessary partnership agreements, with UNHCR’s access to the data and tools. In the future, it will be important for UNHCR to have qualified and dedicat-ed personnel to develop similar approaches to col-lecting, processing, and analyzing big data sources. These efforts should be integrated across different units, to further the Agency’s understanding—as a whole—of the potential of social media and big data to inform operational decisions, advocacy activities, and strategic communications, as well as to improve listening to different affected communities, in order to demystify non-accurate information.

The limitations of the initial interactions and the lessons learned from them helped reshape its initial scope. The project set out to use social media posts to build a better, more nuanced understanding of different complex aspects of the event, for which it was otherwise difficult to assess with traditional tools—such as surveys. However, it now sees added value in trying to use social media to detect unexpected signals of ongoing events that could put PoC at risk, and that UNHCR may need to quickly respond to, or act upon.

The streaming nature of social media posts affords the detection of such signals in near-real-time, which could be useful in cases similar to the aftermath of the Berlin terrorist attack, where more than 17.6 million tweets linked the incident with PoC. There are few data sources that can facilitate such in-depth, rapid response mechanisms, and the project intends to continue exploring their potential.

The WAY FORWARD

UNHCR routinely collects massive amounts of data, through, for example, registration and information management exercises, programme and project im-plementation, and financial activities. The main chal-lenges faced so far, and therefore an area of focus for the Agency, is to find ways of approaching the inte-gration of new data sources into this culture, and to bring more data-driven evidence into decision- mak-ing processes and advocacy efforts, particularly in developing an institutional policy against xenophobia.
African Descent”. UNHCR DIP

Discrimination. Thematic discussion: “Racial discrimination against People of

However, confronting growing intolerance and xe-

nophobia are just some of the many challenges that

may be ahead for UNHCR, in a world that is more con-

nected, and where ideas and words can be shared

across many channels, including digital channels. The

European Network Against Racism (ENAR) published

a study that highlights an increase in protests, political/

elections rhetoric, and formation of structured groups

against refugees and asylum seekers in Europe26. They

mention that “social media is becoming increas-

ingly crucial in forming opinions about migrants, and

there has been a growing dissemination of fake eth-

nicity-related news about migrants with alarming and

sensationalist headlines.”

During the Committee on the Elimination of Racial

Discrimination in March 2011, UNHCR’s Senior Legal

Coordinator explained that “Combating racism, xen-

ophobia and related forms of intolerance against refu-

gees, asylum-seekers and stateless persons is one of

the principle objectives of UNHCR, and these forms

of discrimination are one of the greatest threats to

the rights of refugees and asylum-seekers, in Europe

and elsewhere25. From impacting the right to seek

asylum, to better understanding how xenophobia is

related to the primary root causes of persecution or

negatively affecting integration opportunities, this

is an area of work UNHCR must be more proactive in.

In fact, not addressing xenophobia towards PoC

in a strategic way would constitute a shortcoming of

UNHCR’s overall protection mandate as an agency

The 2009 “Combating Racism, Racial Discrimination,

 Xenophobia and Related Intolerance through a

Strategic Approach” along with the 2015 evaluation

of UNHCR’s Southern Africa Programmes “Protection

from Xenophobia” layout specific guidelines on how

the agency is addressing the issue.

ANNEXES

Annex I: Data Query Taxonomies per Hypothesis

1. O.H Monitor Interactions

- Negative perception: bad conditions in access to services or to territory of asylum, police brutality, closed border, means of transportation;
- Taxonomy: for link: basic neutral, basic posi-

tive, basic negative
- Geography: Greece, national level

Machine learning query: untrained/discard

PoC Farsi:

● Geo: Iran, central level
- obstacle "boat" OR "plane" OR "resettle-

ment"
- OR "boat" OR "plane" OR "relocation" OR "resettle-

ment"

● Geo: Greece, national level
- OR "boat" OR "plane"
- "refugee" AND ("move" OR "movement" OR "move"

- OR "boat" OR "plane" OR "movement"

- OR "boat" OR "plane" OR "resettle-

ment"
- OR "boat" OR "plane" OR "relocation"

- OR "boat" OR "plane"

This last point should further help improve UNHCR’s capac-

ity to make data-driven decisions.

24 Composite data or compound data is any data type which can be constructed in a program using the programming language’s primitive data types. In summary, is any language data type that isn’t a machine number.


Machine-learning query

A) Xenophobia English

(Immigrant OR refugee OR refugees OR immigrants) AND (Greece OR Greeks OR fear OR hatred OR racism OR xenophobia OR foreigners OR arrivals OR Syrians) AND (Immigrant OR refugee OR refugees OR immigrants) AND (RT OR US OR America OR UK OR Trump OR Brexit OR Merkel)

B) Xenophobia Greek

(μετανάστης OR μεταναστής OR μετανάστες OR μετανάστης) AND -(RT OR Βρυξέλλες OR Τσίπρας OR Brexit OR Γερμανία)) OR((μετανάστης OR μετανάστες) AND -(RT OR US OR America OR UK OR Trump OR Brexit OR Merkel))

C) Xenophobia Arabic

(مهاجر OR مهاجرة OR مهاجرين OR مهاجرة) AND -(RT OR Nexus OR Eurozone OR US OR America OR UK OR Trump OR Brexit OR Merkel)

Translation: migrant OR migrants OR refugee OR refugees

D) Xenophobia Farsi

(مهاجر OR مهاجرة OR مهاجرين OR مهاجر) AND -(RT OR Nexus OR Eurozone OR US OR America OR UK OR Trump OR Brexit OR Merkel)

Translation: refugees, refugee, welcome, migrant, Greece, to arrive, foreign, foreigners

1.3 O3

Linking incidents: blame refugees for attacks/incidents, terrorism activities in Europe, Munich, Nice, St. Etienne, #PrayForMunich, #offeneTür, Bastille, #BerlinAttack

Taxonomy: blame refugees, do not blame refugee, neutral, irrelevant

Geography: Worldwide

Machine-learning query

A) Situation Awareness Munich

(Munich OR MunichAttack OR PrayForMunich OR offeneTür OR Beschuldige OR Fluchttügel OR Flüchtlingen OR Schuld OR Attacke OR Tod OR Toten OR Opfer OR Schießen OR Schiessen OR Attentäter OR Gewehr OR Pistole J OR [attack OR killer OR killed OR dead OR deadly OR death OR shooting OR gun OR bullets OR victims OR killing]) OR (München OR München OR Мишен OR Мюнхен OR מינכן OR ميونخ OR Μόναχο OR Μόναχο OR Munich OR PrayForMunich OR NavParams OR κατηγορούν OR κατηγορία OR κατηγορία OR ισλαμοφοβία OR θάνατο OR τραυμα OR θυσία OR ρατσισμός OR θάνατο OR θήραμα OR θηραμα OR ισλαμοφόβια OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλαμοφοβία OR ισλα}
Annex II: Tweets found and catalogued by AI

O1H: Monitor Interactions

Translation: You are frustrated by all the refugees dying in the sea but words don’t do us much, open the borders

<table>
<thead>
<tr>
<th>Source</th>
<th>Post</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>October 30, 2015</td>
</tr>
</tbody>
</table>

Translation: The governor of Greek Central Macedonia: There are about 13,000 refugees are swarming to the Greek-Macedonian borders in miserable conditions

<table>
<thead>
<tr>
<th>Source</th>
<th>Post</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>March 6, 2016</td>
</tr>
</tbody>
</table>

Translation: Greece is currently facing a huge economic crisis. And the circumstances for the refugees are even more difficult

O2H: Understanding sentiment

Xenophobia English Monitor
Category: Xenophobic

Neutral

Irrelevant
On Monday the first Syrian refugees will move from Turkey to Germany.

Annex III: Data Visualizations (Quantitative inputs)

O3H: Incidents Linkage
Total Number of tweets analyzed: 3,433,800 (Nice) + 297,506,445 (Munich) = 300,940,245 posts up to Jan 10th, 2017.
Xenophobic: Munich (8%) in yellow and Nice (6%) in purple
Geography: Worldwide

Xenophobic

Non-Xenophobic

Neutral

Nice and Munich (January 10th, 2017)

O3H: Incidents Linkage
Total Number of tweets analyzed: Munich (58,815,918 posts), Nice (3,748,198 posts) and Berlin (353,580,956 posts) = total 416,435,072 posts
Xenophobic: Munich (8%) in yellow, Nice (7%) in green and Berlin (5%) in purple
Geography: Worldwide
Xenophobic
Not-Xenophobic

**Hypothesis:** Incidents Linkage

**Total Number of tweets analyzed:** Munich (58,815,918 posts), Nice (3,748,198 posts) and Berlin (353,580,956 posts) = total 416,145,072 posts

**Non-Xenophobic:** Munich (<1%) in yellow, Nice (11%) in green and Berlin (7%) in purple

**Geography:** Worldwide

---

**Annex III: Data Visualizations (Qualitative inputs)**

*Data Visualization type: Word Cloud, Munich situation awareness monitor*

**Data Visualization type: Cluster, Munich situation awareness monitor**
Annex IV: Interactive map
(under-construction by UNGP)

Tweets geo-located by route, interactive map (under construction). Python based
How to cite this document:

The opinions expressed in this paper are those of the authors and do not necessarily represent the position of UN Global Pulse or UNHCR.