Empowered lives.
Resilient nations.

Digital Lighthouse Initiative

Applying Big Data and AI in the context of Hate Speech across Social Media

Regional Bureau for Arab States

Future Forward

UNDP Digital Strategy

Table of Contents

I. SITUATION ANALYSIS / BACKGROUND

Countries of the Arab region continue to face threats to basic human rights, social cohesion, diversity, security, justice and tolerance. The Arab region is complex yet rich, within which exists a social and cultural diversity that manifests itself in multiple dimensions: ethnic, tribal, religious, and linguistic, resulting from overlapping factors that relate to history, tradition, and immigration. This social and cultural diversity is facing increasing threats due to various internal and external factors, be it socio-political, economic, environmental and geopolitical. The weakening of tolerance and acceptance of the "other" is manifesting itself in rising hate speech, especially on social media platforms. At the same time, freedom of speech and expression has been witnessing a setback in some countries in recent years, despite resistance from citizens who are becoming more vocal in voicing their demands.

Through its various initiatives at the country and regional levels, UNDP RBAS is committed to reverse the negative trends and address drivers of conflict, intolerance and polarization in this region, in line with UNDP's Strategic Plan, contributing mainly to Signature Solution 2 "strengthen effective, inclusive and accountable governance" and Signature Solution 3 "enhance national prevention and recovery capacities for resilience societies". But producing relevant and timely analysis and programmatic recommendations is proving to be both difficult and expensive especially in conflict and conflict-affected countries where traditional mechanisms of data collection are either disrupted or manipulated.

To understand complex trends in people's perceptions, attitudes and actions, big data and AI technologies carry the potential to be as relevant, timelier and more cost effective than the traditional data collection mechanism and "could make the data cycle match the decision (and programmatic) cycle"[1]. Many big data providers crawl information from more than 150 Million sources worldwide, including data collected from websites and social media (twitter, Forums, blogs, etc..), online and offline media (TV/Radio, prints and images). These providers offer access to datasets through platforms powered by advanced search capabilities and filters in more than 18 languages (including the 6 main UN languages) and benefit from AI modules to guaranty accuracy of results. Big data technologies are being increasingly leveraged in the development and humanitarian sectors. However, given its "perceived" complexity, lack of resources, political sensitivities, and limited understanding of its potential, the peacebuilding arena has been more reluctant to capitalize on these promising technologies.

With the aim to capitalize on the potential of big data and AI technologies, this Project will explore the use of these under-utilized and promising technologies - without compromising impartiality, privacy rights, and methodological soundness – to monitor the rise of hate speech in a world that is being shaped by new developments in internet and communications technology (ICT) and rapidly increasing mobile phone and internet penetration, making the availability and volume of data expand at an exponential rate.

In the Arab region, mobile-cellular subscriptions have increased by around 40% between 2010 and 2018 and estimated to have reached a penetration rate of 103% by the end of 2018. The percentage of individuals (% of total population) using the internet in the Arab region is also expected to have more than doubled by the end of 2018 at 54.7% up from 24% in 2010.[2] With the proliferation of ICT access, social media has grown into a powerful tool of political and social interaction and opened the door for self-expression. The total number of monthly active Twitter users in the Arab region is estimated to have reached 11.1 million in March 2017, up from 5.8 million in 2014, of which 29% are in Saudi Arabia and 18% in Egypt. Out of the 166 million internet users in 13 Arab countries, 66% (or 110 million) have opened Facebook accounts by December 2017.[3]

1 "A World that Counts" available at http://www.undatarevolution.org/wp-content/uploads/2014/11/AWorld-That-Counts.pdf

2 https://www.itu.int

3 https://www.internetworldstats.com/stats5.htm accessed in May 2019.

ICT has also in one-way incentivized violence by offering a relatively cheap space to attract audience. An open space of freedom of expression has also indirectly led to the spread of hate speech and discriminatory language and attitudes has become amplified, especially against groups that are an easy target even outside virtual space. In one way, the virtual world has also reaffirmed hierarchical and discriminatory attitudes towards certain groups like women, LGBT, ethnic/religious minorities, refugees and migrant workers. In certain cases, the virtual space has mimicked the power relations that exist in the real world, and that are systematically maintained by the de facto legal and institutional framework that does not treat all citizens equally. As such, the move from the non-virtual to the virtual world does not mean that the less privileged groups will live in a safe haven. There is of course an increased opportunity for self-expression, but it does not always come without a cost. Arrests over social media posts has recently become the norm rather than the exception in Lebanon, Egypt, Jordan, Saudi Arabia and other Arab countries.

Monitoring hate speech in the Arab region should take into consideration deteriorating levels of freedom of expression on specific controversial issues coupled with an increased space for freedom of expression for certain groups at the expense of others made possible by more accessible social media platforms. This Project will benefit from previous work conducted by the UN Global Pulse such as “The Effect of Extremist Violence on Hateful Speech Online” (April 2018) to refine the research methodology, but it will propose a different conceptual framework, whereby reactionary verbal violence and spikes of hate speech will be acknowledged but filtered out in an attempt to draw a mapping of systematic, deep-rooted and entrenched hate speech (drivers, influencers, framing, etc.) in the societies. The other side of the coin is to form an understanding of the context when hate speech content is minimal and when the online community shows interest to discuss other matters, that are perhaps non-divisionary.

Now the big elephant in the room is how to classify content as hate speech, and this calls for the adoption of a functional definition of hate speech. As such, the initiation phase of this Project would necessitate conducting a series of focus groups with human rights experts, global and national, to agree on a definition that is in line with the Universal Declaration of Human Rights, related conventions, and the legislative infrastructure in the respective countries.

Country selection would also help in enriching the conceptual framework, especially with reference to drivers/type of hate speech and inter-group dynamics. To enrich the design of the research tool, the inception phase will look into the case of:

- **Lebanon**: multi-religious/sect country with a big migrant community and among the highest refugee per capita rate globally.
- **Egypt**: home to around 20% of Twitter account, a big Copt/Christian community (9-10 million) in a Muslim-majority country
- **Tunisia**: homogenous country but ideological polarization among Islamists and secularists is on the rise.
- **Or Iraq**: multi-ethnic/religious country with high polarization among political parties that are relatively new.

II. VALUE PROPOSITION AND TARGET USERS

Value Proposition

This Project will develop a rights-based conceptual framework and design a research tool centered around the use of big data to build a deeper understanding of why, when and how hate speech spikes (what prompts someone or a community to express themselves this way, what reactions are there to it (to what extent is it endorsed and by what profiles, to what extent is it repudiated and by what profiles), what is the role of networks and collective action in amplifying or shutting down hate speech, etc.) Similar attempts by UNDP Country Offices (Cos) did not take the full leap forward, and while social media content has been used by a few COs, the analysis was based on manual data collection rather than an AI and machine learning modality that allows for reaping the benefit of the full value-chain of big data: descriptive, diagnostic, predictive and prescriptive analytics. It will help establish UNDP as an early adopter of big data technology and help in upskilling and retraining project managers and coordinators in new areas.

For the wider community, this Project will offer a multi-disciplinary and multi-lens methodology that helps in building a deeper understanding of drivers of hate speech as an expression/aggression that can be systematically practiced against a certain group(s). For example, the UN Global Pulse has published in April 2018 an important study titled “The Effect of Extremist Violence on Hateful Speech Online”, which basically monitored the reaction of the online community towards an incident of violence, thus focusing on spikes of hate speech, rather than analyzing disguised hate speech during times of (relative) peace. This Project will build on the four-dimensional framework of classifying hate speech and counter- hate speech content; (1) stance, (2) target), (3) severity and (4) framing, but it will try to look for the day-to-day hate speech content that is even more deeply entrenched in societies, especially against groups that are affected by the imbalance in power dynamics at the level of the state and its institutions (security, judiciary, representation, etc.) whereby certain groups are given more freedom and more rights compared to others.

Fortunately, machine learning tools can identify when hate speech is occurring, who is interacting with hate speech content and where are influencers or those who initiate hate speech content are located. Data from ICT sources such as text and image can be collected from social media and the internet. This includes social media scraping (Twitter, Online News, Instagram, Blogs, Newspaper, Forums, Flickr, Youtube, etc..) , offline and online media analysis (TV, radio with voice recognition technologies, prints, etc..).

This Project will invest in accessing big data from pre-existing (datasets that offers not only access to archives (period can span from 4 to 2 years) but also collects news data and social media data (Twitter, Facebook, etc.) and gather data in real time to produce structured information on the levels and source of hate speech, feelings of exclusion and attitudes of intolerance in the Arab region.

It will support multidisciplinary collaboration and research and nurture innovative research ecosystem to address the Arab region’s most pressing challenges to safeguard diversity by bringing together data scientists, sociologists, development practitioners, statisticians and others.

Here, it is important to clarify that machine learning does not exclude human expertise, knowledge and contribution to the design of the framework and analysis of results. Machine learning can identify patterns, but it still takes a human eye to understand and contextualize those patterns in support of policies to reverse the trends.

In designing the conceptual framework, the Project will reach out to similar research projects that have been conducted recently by leading academic institutions (Columbia University, University of Cardiff, etc.). It will also try to build but expand on the work initiated by the UN Global Pulse and position Arabic as one of the pilot languages of the SG Working Group on Hate Speech, to which it was mandated the task of developing language models around hate speech. Hence, a close collaboration with the Global Pulse team has been initiated.

Moreover, the framework and big data queries will have to be contextualized to the Arab region and selected Arab countries by working with a group of experts from these countries to confirm veracity,

nuance, intergroup dynamics, etc. Here, it is important to take into consideration the risk of designing a system that can be abused by security institutions to silence free speech and free expression.

The project aims to support the region's most creative and forward-thinking researchers from all disciplines to work together and explore how to design a query-based and algorithmic framework that can help simplify complex discriminatory and hate relationships among various and multi-layered groups in the society.

Target Users

For whom is it solving problems?

The Project will create two levels of value proposition:

(1) External targeting development and human rights practitioners and academics, policy makers, civil society organizations, in addition to other agencies who have either attempted to conduct similar work by helping them refine the research tool or improve their programming and delivery.

   a. Increase synergies and collaboration between research and innovation projects in the Arab region by developing an integrated model that brings together machine learning with a multi-disciplinary group of researchers, data scientists and development practitioners;

   b. Position UNDP RBAS as an early explorer in future technologies in the Arab region, especially in the field of peacebuilding.

(2) Internal targeting UNDP programming

   a. Produce timely and relevant information on intensity, source and drivers of deep-rooted rather than reactionary hate speech that is triggered by an incident of violence;

   b. Modernize system(s) of data collection and reframe how big data is viewed;

   c. Develop a visualization and interactive system in the form of a dashboard and processes that can be easily replicated by other regions and country offices at UNDP;

It will support multidisciplinary collaboration and research and nurture innovative research ecosystem to address the Arab region's most pressing challenges to safeguard diversity by bringing together data scientists, sociologists, development practitioners, policy makers, statisticians and others.

III. OBJECTIVES AND KEY RESULT

Objectives

There are three sets of objectives that this DLI will help achieve: long-term strategic objectives, medium-term objectives and immediate/short-term objectives.

In the long-term, this Project will:

- Position UNDP as an early adopter of new technologies to support peacebuilding in the Arab region;
- Support the design of more targeted programmes to counter hate speech in the Arab region; and
- Push for more inclusive non-discriminatory legislation based on international human rights laws.

In the medium-term, this Project will:

- Reconcile various disciplines with big data as a complementary source and mechanism of validation of perceptions and attitudes on various topics, taking into account limitations and privacy rights issues.
- Offer advocacy and human rights groups the resources and tools to improve their campaign;
- Stimulate the rise of new research areas and businesses in the Arab region.

In the short-term, this Project will:

- Build a business model that can be replicated by other UNDP colleagues to dig into other topics;
- Offer a research methodology and rolling-out model engined by internal know-how (within UNDP) to help upskill and retrain project managers and all those relevant to help them identify new areas of research and programming that were not possible or too expensive in the near future;
- Deepen the understanding on hate speech by producing timely sentiment reports in a region that has become home to the highest number of conflicts, refugees and IDPs in recent history.
- Contribute to refining existing research tools on similar topics.

Key Result

The achievement of the short-term objectives mentioned above will require the DLI deliver on the following major key results by the end of 2019:

(1) Screen at least five external data providers and identify one suitable supplier for the DLI exercise to be conducted

(2) Provide data access to Social Media Data suitable for the 3 Arab countries in scope

(3) Recruit 25 expert participants from 3 Arab countries to participate in the hate speech discourse and

(4) Conduct 3 dialogue consultation sessions to validate the framework and methodology

(5) Develop hate speech taxonomy and construct query model for 3 Arab Countries

(6) Develop one proof of concept dashboard interface with the relevant criteria for defined potential user to be used for internal as well as external

(7) Create one evaluation report with 3 contextual action areas for each of the 3 Arab countries

IV. DELIVERABLES

The work of the Digital Lighthouse initiative has been split into six work packages which define the scope and deliverables of this project.

Work Package 1 – Digital Lighthouse Initiative Setup

- Defines the project objectives, timeline and team setup (incl. governance) to start the project. Also gather the necessary resources (internal as well as external) for the creation of a multidisciplinary team including Data, ML, political and sociologist experts, define TOR if necessary and identify & include any external partners for the development of the hate speech monitoring solution. Also include the planned budget estimation.

Deliverables: Project documentation (incl. budget) and taskforce in place

Work Package 2 – Data Access

- Identify the necessary data sources required, identify the data vendor/provider and receive access to the data. The process of identification involves a screening of current UN initiatives which source social media data already and conducting a data vendor selection process relevant to the requirements of the contextual analysis.
- This process might also include exploring the option of engaging the services of a ready social media monitoring platform, a one-stop-shop data provider, which grants access to historical and live data from multiple social media and other online/offline media sources (e.g. news, blogs, etc.).
- The preliminary research conducted by the UNDP team has helped in developing a base of selection criteria to help in screening data providers (see Annex: Vendor Selection Criteria (Draft) for detail)
- An external data scientist will provide important input to assist in selecting the data provider, after reviewing other UN initiatives: whether selecting between accessing the data directly from the social media api (e.g. twitter, reddit, etc.) and build our own UNDP search interface or engaging with one-stop-shop providers which hosts the data and provides advanced filtering and AI capabilities. The latter is preferable given the time frame of this initiative, but at a later stage, UNDP should invest in designing such platforms for cost-efficiency reasons and to build internal capacity to allow UNDP to become a fast-adapter.

Deliverables: Requirements for the social media data identified, vendor selection completed and data sources & access acquired

Work Package 3 – Contextualization of hate speech

- The research team will outline the definition of hate speech based on given conventions and build on existing research conducted by the UN network (UN Global Pulse).
- Furthermore, the team will explore other given frameworks, experiences and case studies to help develop a better understanding in shaping the requirements for the contextualization of hate speech.
- In addition, the research team will develop a framework and methodology (incl. a glossary of hate speech) to evaluate hate speech in selected Arab countries - scope will cover 3 Arab countries in 2019 (Lebanon, Iraq/Tunisia and Egypt).
- The framework will include research with social media data to explore areas around hate speech in connection to other social economic and political indicators.

Deliverables: Curated research on the topic of hate speech in social media, framework and taxonomy on hate speech in place and tested in workshop with expert representatives

Work Package 4 – Evaluation of results

- The research team together with the programme specialist team will explore initial results and findings from applying the methodology and framework developed on a large set of social media data.
- Discuss preliminary findings and agree on validation mechanism with internal team of the UNPD and extended research group to identify major fields of action for the relevant countries within their specific context
- A sentiment report that offers a summary of preliminary findings and three top fields of action based on initial findings from analysis (disaggregated by governorate at country level, country, social group, age group, etc.) will be developed

Deliverables: Sentiment report on hate speech in the region – focus on three countries

Work Package 5 - Dashboard

- The project team will develop an internal/restricted access dashboard showcasing findings of selected visualizing the results of the research Arab countries that show intensity, driver and dynamics of hate speech among various groups, identifying aggressors and victims of hate speech.
- Dashboard will include user-friendly data visualization charts and maps to present different query results and analysis from big data. On technology side, UNDP team will explore the usage of available UNDP systems licenses (such as Power BI) combined with other programing languages to build the interface.
- Part of the DLI until end of 2019 will be the launch and test of an internal proof of concept of the dashboard, which will potentially be launched in Spring 2020 - after successful feedback – to a wider audience for policy makers and possibly the public.

Deliverables: Dashboard with key features for visualization, filters, interactive map & analysis.

Work Package 6 – Dissemination and Communication

- With the preliminary results and the dashboard in place a roadmap will be prepared for rolling out and disseminating (and translating) the methodology alongside capacity building to support its implementation of multidisciplinary teams on the evaluation and usage of Big Data and AI for better programming.
- Also, the work package will help identify long-term owners, develop a change management plan and conduct dissemination event to offer advice to position UNDP as thought leader in the space of analyzing social media data in the context of better programming.
- 3 Expert Group and Dissemination Meetings to discuss preliminary findings and agree on validation mechanism developed by the research team
- Conduct soft/internal launch of preliminary findings with potential donors, UNDP partners to make results of the research available to public – if relevant permission is provided (considering data security and legal aspects regarding free speech)

Deliverables: Soft launch event, communication, (Beirut or Tunisia – TBC,) Expert Group meetings

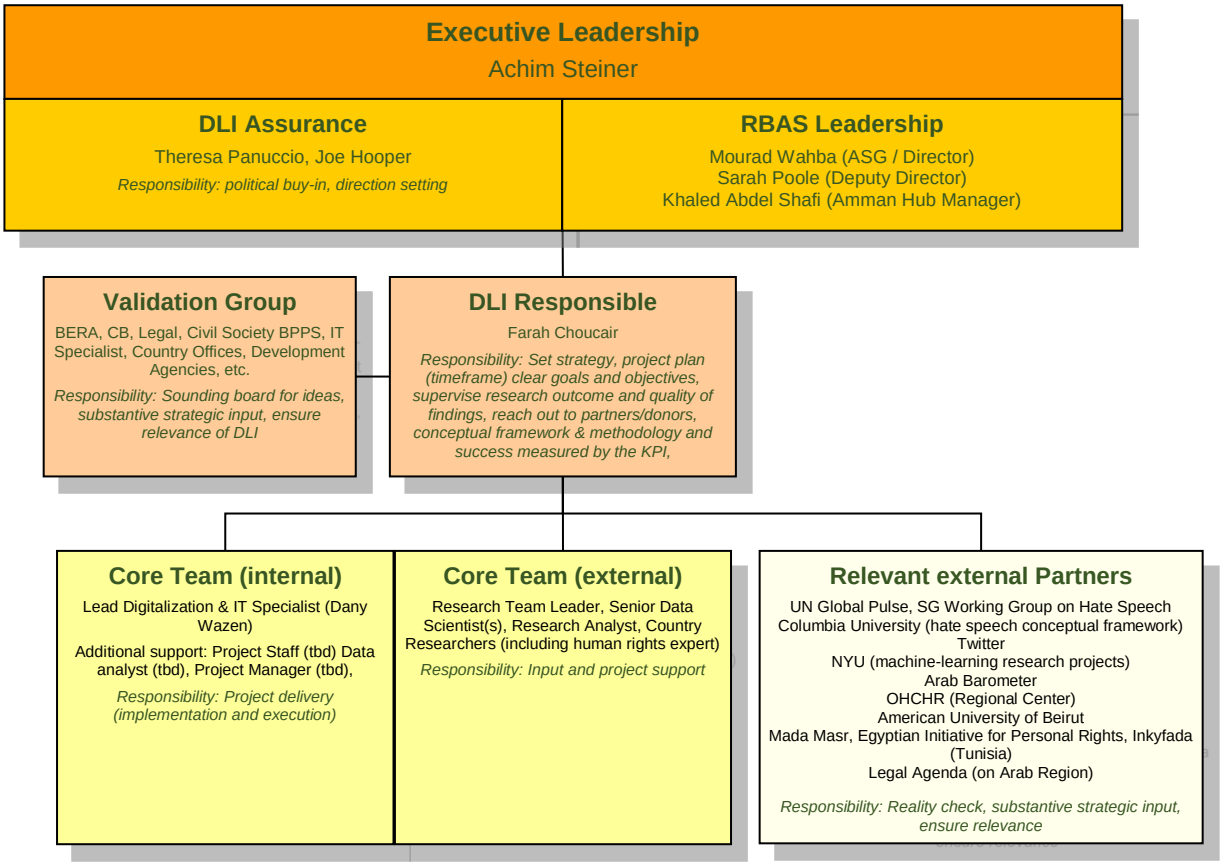
V. WORK PLAN

What are critical milestones and respective activities?
Who is responsible for carrying out the activities and what resources do they require?

Work packages *Add work packages, including high level result*	**PLANNED ACTIVITIES** *List activity results and associated actions*	**TIMEFRAME**								**RESPONSIBLE PARTY**
		May	Jun	Jul	Aug	Sept	Oct	Nov	Dec	
Work Package 1 *Digital Lighthouse Initiative Setup*	Create internal taskforce (Project manager and UNDP team)	*End*								*Farah Choucair*
	Project documentation (incl. timeline, resources, responsible and budget) completed	*End*								*Farah Choucair*
	Receive internal leadership and budget approval		*Mid*							*Farah Choucair*
	Onboard additional external research partners for the initiative – free of cost (Academic institutions, private partners or other development institutions)			*Mid*						*Farah Choucair* *Jos De La Haye, Malin Herwig*
	Procure and onboard additional external resources to taskforce – cost related (Define TORs, conduct hiring process – senior researcher, senior data scientist, research/project assistant, country researchers)				*Mid*					*Farah Choucair* *Jennifer Colville* *Dany Wazen*
Work Package 2 *Data Access*	Agree on the privacy and ethics disclaimer and requirements		*End*							*Farah Choucair*
	Identify potential data provider with the necessary access to required data points across social media platforms and thematic keywords search		*Mid*							*Dany Wazen*
	Screen potential providers for given agreements with the UN network		*Mid*							*Dany Wazen*
	Define data requirements needed for the AI-engine dashboard and disaggregated outcomes		*End*							*Farah Choucair, Paola Pagliani, Dany Wazen*
	Develop data hosting concept for the DLI (hosting of social media data, hosting of query and analysis capacity)		*End*							*Dany Wazen*
	Create data provider criteria list and questionnaire for selection process (e.g. total number of allowed results and pricing, query limit, given analysis tools, ai module, historical data options, export capabilities, etc.)			*End*						*Farah Choucair, Paola Pagliani, Dany Wazen*
	Conduct data vendor selection process					*End*				*Dany Wazen, Maya Baydoun*
	Procure access to data provider (contractual agreement)						*Beg*			*Maya Baydoun*
Work Package 3 *Contextualization of hate speech*	Agree on a functional definition of hate speech within the framework of the Universal Declaration of Human Rights and other relevant conventions			*End*						*Farah Choucair and Marta Vallejo*
	Screen and align with existing projects on hate speech contextualization on social media (internally/externally)			*End*						*Farah Choucair, Paola Pgliani*
	Initiate dialogue to collect additional experiences, case studies and feedback on the approach (Identify stakeholders, matter experts and define dialogue question)					*Beg*				*Farah Choucair*

	Draft research methodology and initial structure of contextualization framework and taxonomy for hate speech to be discussed with expert research group in workshop format					End				*Farah Choucair*
	Prepare, plan and conduct research workshop meeting with experts and representatives from the respective countries (context of 3 Arab countries)					End				*Farah Choucair*
	Evaluate feedback from workshop and refine contextualization framework and taxonomy by producing country mapping (key words, timeline of hate speech and major triggering events, identification of groups)							End		*Research Team Leader*
	Translate framework into technical requirements to be applied on historic, live and for-ward looking data (Machine Learning)								Beg	*Research Team Leader*
	Revisit research methodology and taxonomy								Mid	*Research Team Leader*
Work Package 4 *Evaluation of results*	Evaluate data points based on contextual framework and country profiling of major socio-economic and political developments/trends						End			*Research Team Leader*
	Identify three top fields of action based on initial findings from analysis							Beg		*Research Team Leader*
	Discuss preliminary findings and agree on validation mechanism with internal team of the UNPD and research group							Mid		*Research Team*
	Generate summary report of preliminary findings and policy-level implications to trigger discussion								Beg	*Research Team Leader, Farah Choucair, Paola Pagliani*
	Translating and printing report (Completion: Spring 2020)								End	*Farah Choucair*
Work Package 5 *Dashboard (proof of concept – internal launch by end of 2019)*	Detail proof of concept of the design of the external dashboard (incl. requirements)			End						*Dany Wazen Pierre Hamouche*
	Identify technical solution (incl. technology and hosting option) for the dashboard				End					*Dany Wazen Pierre Hamouche*
	Develop user interface for the dashboard and available charts (Pilot for testing)							Mid		*Dany Wazen Pierre Hamouche*
	Connect pilot dashboard with database							End		*Dany Wazen Pierre Hamouche*
	Test proof of concept with expert group from representative countries								Mid	*Farah Choucair*
Work Package 6 *Dissemination and Communication*	Prepare, plan and conduct expert group dissemination and consultation/validation meetings in each country to discuss preliminary findings	*Start at the end of 2019, actual activities planned for spring 2020*								*Farah Choucair & Team*
	Prepare, plan and conduct (a soft) launch event of preliminary findings (internal with selected external participation)									*Farah Choucair*
	Manage stakeholder and involved parties									*Jennifer Colville*

VI. MANAGEMENT ARRANGEMENTS



Other resources

The Digital Lighthouse Initiative will rely heavily on the dataset which will purchased from a suitable data provider which has the right data sets and also the right features to apply analysis.

Partnerships (resources)

Partnerships and fellow researcher are being leveraged as part of the Digital Lighthouse Initiative. This is key in ensuring that existing frameworks, research and models are being used in this project to achieve faster and better results. Existing initiatives and potentials partners which are currently conducting hate speech initiatives include:

- United Nations Global Pulse (Link)

  To link initiative to the SG Action Group on Hate Speech and offer technical assistance from their senior data scientist(s) who was part of a similar initiative plus additional advice on retrieving large datasets from platforms such as Twitter.
- Hatebase.org – Qatar Research Institute (Link)
- Columbia University – Initiatives on Hate Speech and use of Big Data on Social Media

  Benefit from their conceptual model to monitor hate speech.
- Anti Defamation League (Link)
- Cross Cut with economic institutes and social aspects
- Office of National Statistics (Link) – Finding insights with breaching data privacy

VII. BUDGET PLAN

PLANNED BUDGET		
BUDGET ITEM	**NOTES**	**AMOUNT (USD)**
Digital Lighthouse Initiative Setup	*Subtotal*	*44,500*
*Procurement and project support cost recovery***	*5% rate of procurement contingency fee (25,000 USD) + project support related fees (5,550 USD)*	*30,550*
Travel	*Project related contingency for travel cost related to the work of the DLI*	*14,000*
Contextualization of hate speech & evaluation	*Subtotal*	*142,250*
Research Team Leader	*1 Person (60 days of consultancy services) Lead researcher will develop the conceptual framework, methodology, query construction, conduct analysis and produce sentiment report*	*45,000*
Country Researchers	*9 Persons (3 researchers per country) Research consortium will develop the country glossary and identify groups (victims and perpetrators). They will engage in a detailed discussion and react to data findings.*	*45,000*
Data scientist (Arabic Speaker)	*1 Person (40 days of consultancy services) Data scientist specialist in social media data, big data and machine learning*	*20,000*
Data scientist (UN Global Pulse)	*1 Person (15 days of consultancy services Will support the data scientist and share lessons learned from past experiences*	*11,250*
1 Research Assistant (full-time, 6-month period)	*Research assistant will conduct country mapping, prepare country profiles, follow-up with country researchers, prepare summary reports, etc.*	*21,000*
*Data Access**	*Subtotal*	*98,200*
Data provider	*3mn users per month credit streaming (forward for 12months) including twitter and other websites data including 1M twitter data*	*55,000*
Additional 1M streaming forward cost	*Additional 1mn (on top of the 3mn) for forward streaming (over the period of 12 months)*	*14,400*
Adding 1M backward	*100 months of streaming backward to get 1mn/month*	*28,800*
Development of dashboard	*Subtotal*	*22,000*
Web developer	*The total work estimated as per the timeline is around 77 days (within 3 months). UNDP RBAS Hub has an LTA signed with a developer. The developer rate is 23.5USD per hour. The total amount will be around 15,000 USD*	*15,000*
Design	*1 Person (30 days of consultancy services) UNDP RBAS Hub has an LTA signed with a designer. The designer rate is 100USD per day. The total amount will be around 3,000 USD*	*3,000*
Hosting	*Average acceptable hosting as dedicated server is around 300 USD per month (3,600USD per year) in addition to the cost of domain name and certificate is around 400 USD a year.*	*4,000*
Expert Groups and Dissemination	*Subtotal*	*193,000*

3 Expert Groups and Dissemination (EGMs)	*A meeting in each country to discuss preliminary findings and agree on validation mechanism in the country (policymakers, CSOs, researchers, government, human rights institutions, etc.)*	*36,000*
Joint workshop between country teams to discuss taxonomy and finalize country mapping	*Travel and accommodation of 15 participants for a 3-day workshop in Beirut*	*18,000*
Soft/internal launching of preliminary findings	*bringing together potential partners, donors, UNDP COs/RRs, policy-makers, academics to debrief on findings in 1 day event NYC (travel and accommodation of 25 participants from Europe and Arab countries)* *2 nights of stay in accommodation = 20,000 USD* *25 x 3,000 USD flights = 75,000 USD* *Venue rent & promotional material and Running cost of the event (incl. translators) = 25,000 USD*	*120,000*
Translation and communication products (spot, summary report, etc.)		*19,000*
TOTAL		***500,000***

** Budget allocated to data provider is based on exchange with https://www.crowdanalyzer.com/, which is considered the leading Arabic Social Media Monitoring Platform. The cost is at the upper range and might be lowered if the current agreement between Global Pulse and Twitter is extended to UN agencies to allow for free access to Twitter accounts.*

*** Staff recovery cost including efforts from DLI Project Manager, Digitalization Lead Expert, Innovation Team Leader and Governance Team Leader will be waived and covered by the bureau.*

VIII. ANNEX

Target Users (Detailed)

Persona 1 (external)

Role: Private Sector (ex. Facebook)

Objectives: Acquire data, knowledge, funding / Influence leverage / CSR / Include as a partner early on, i.e. National Dialogue

Persona 2 (external)

Role: Policy Maker

Objectives: Create national action plan to address hate speech / Create better policies to address hate speech / Guide national dialogue to address hate speech / Better understand hate speech in their country / A response beyond criminalization

Persona 3 (external)

Role: Academia

Objectives: Better understanding of social fabric / Contribute as partner on socially rewarding projects / Understand potential usage of new technology / Influence education and next generations

Persona 4 (external)

Role: Social Media Users (Youssef and Karima)

Objectives: Source of information and identify impact of language / Create social norms around hate speech / empower victims of hate speech / Guidance on response to hate speech

Persona 5 (external)

Role: Media Journalist – Annahar (40+)

Objectives: News coverage / Analysis & fact validation / Promote values subjectively / Enhance media policies (ex. Facebook)

Persona 6 (internal)

Role: UNDP COs Rep OR Country Team Specialist – Maya (30+)

Objectives: Implement UN Strategy / Early warning and prevention strategies / Support policy making / Support and development of legal frameworks and implementation support / Advocacy on human rights and women empowerment / Insights, analytics, understanding trends, capture lessons learnt for adaption or replication / support adaptive management (UNDP – other agencies)

Persona 7 (external)

Role: Politician - Pierre (50)

Objectives: Promoting political interests and objectives / self-evaluation / Better understand limits of political narratives / debate

Persona 8 (external)

Role: CSO – Love & Peace (50)

Objectives: Positive advocacy campaigns / Adaptive Management to an event / Support drafting on legal framework / Partner with UNDP to implement solutions

Key Features (Detailed)

In scope for the first six months

Dashboard

- Data visualization
- Allow for filter function to apply big queries and the usage of machine learning
- Application Programming Interface (API) through the Datasets – interfacing between the dataset providers and the platform
- Dashboard country profile (historic) – interactive map to display data

Hate speech contextualization

- Algorithm design – Developing an algorithm to identify and monitor hate speech
- Contextualization of hate speech – Taxonomy by country / Core – Multi disciplinary research team
- Validation function (mechanism and processes) – validate / discuss findings / analysis, feed back into design of algorithm by research team, country offices, media and political parties

Additional Key Feature relevant after six months

Dashboard

- Real time alert (country profile) – Hate speech alert, possibly covered by crisis dashboard
- Linkages to other knowledge assets, e.g. Crisis Risk Dashboard
- Analytics feature
- Monitoring – to what extend is our programming having an impact on hate speech and what are the results

Expertise

- Implications for policy and programmes – early warning mechanism
- Co-creation of solutions feature – Discussion for community of practice
- New Partnership and Business Model – Mapping, who does what in Private sector, civil society (UN Hate Speech Agenda, UNDP)
- Mapping related projects (externally) in the field of hate speech monitoring
- Risk Management – Understand, manage and mitigate the risk associated with working on big data an AI

Risk Analysis

CONSTRAINT (INTERNAL) / RISK (EXTERNAL)	DESCRIPTION	MITIGATION / STRATEGY
Risk (external)	Reliability and integrity of big data sets *Probability: Medium* *Impact: High*	1. *Selection of known and recognized data provider. (Check with global market research companies)* 2. *Clarify scope and limitations of datasets used.* 3. *Onboard data scientist to help shape the requirements for selection*
Risk (external)	Application of non-relevant conceptual framework and Big Queries *Probability: High* *Impact: High*	1. *Reach out (and partner) to similar research projects by leading academic institutions for assistance* 2. *Validate framework with national/local experts*
Risk (external)	Varieties of Arabic language dialects *Probability: Medium* *Impact: Medium*	1. *Application of ML to refine the query results.* 2. *Creative and forward-thinking researchers from all disciplines to work together and identify nuances*
Risk (external)	Reliability of AI and Machine Leaning capability *Probability: Medium* *Impact: High*	1. *Selection of known and recognized AI experts (check within partner network on recommendation)* 2. *Clarify scope and limitations of technology used.*
Risk (external)	Reputational risk to engage on hate speech discussions *Probability: Low* *Impact: High*	1. *Involve legal experts to mitigate any grounds for misuse constitutional violation* 2. *Chose countries*
.Risk (external)	Civil lawsuit filed by a hate group against UNDP citing ‘free speech’ based on constitutional grounds *Probability: Low* *Impact: High*	1. *Consultation of external experts to build capacity through experience internally*
Constraints (internal)	Limited expertise on Data / AI / Hate Speech *Probability: High* *Impact: High*	1. *Consultation of external experts to build capacity through experience internally*
Constraints (internal)	Project is not covering all countries in the Region *Probability: High* *Impact: Low*	1. *Identify additional countries to be launched after successful pilot* 2. *Communicate launch plan for additional countries in the region*
Constraints (internal)	Long procurement process might challenge timeline *Probability: High* *Impact: High*	1. *Mitigate long procurement process through involvement of executive leadership*

Cost Effectiveness Analysis (CEA)

The cost effectiveness analysis (CEA) of using the big data and AI approach seeks to validate the decision to opt for the best alternative activity, process, or intervention that minimizes resource use to achieve a desired result.

UNDP Country Offices continues to invest in high-cost nationally representative surveys, which are a necessity in certain cases, such as conducting household surveys to measure poverty and livelihoods conditions or labor surveys to produce labour statistics. Discussions with leading surveying firms to run nationally representative surveys can range from USD 50,000 (for a sample of 1,400) up to USD 150,000, noting that nationally representative surveys are not representative enough to produce governorate-level analysis. However, in other cases, big data offer a better medium to measure perceptions, attitudes, feelings and collective action towards issues that are considered sensitive. Inter-group dynamics is one of those areas that can be better captures using big data and social media platforms, that act as a space of unfiltered self-expression.

The usage of big data technology can yield more representative findings for a larger number of users at a lower cost. The cost effectiveness will be achieved through big data and AI to fill the missing gaps in the traditional surveys. Not using big data and AI will be a missed opportunity to identify alarming trends and foresee incidents of violence and/or identity-based polarization at a lower cost and with greater replication effect for other countries. Additional, traditional mechanisms of data collection might not be an option for countries in violent conflict or polarized societies, where the security of both data collector and respondent might be at risk.

The budget of USD500,000 will be allocated across four main components: (1) Formation of a country-level research consortium and a multi-disciplinary team to develop the conceptual framework and contextualize hate speech taxonomy and construct query model, (2) Access big data on hate speech through a data provider (here, it is important to mention that the cost would drop significantly if the UN reaches a legal agreement with Twitter granting it free access to its accounts, (3) development of an dashboard to offer access to aggregated findings on countries and hate speech content (internal access at first and (4) conduct focus group meetings and a general soft launching event to discuss preliminary findings with potential donors and partners (opportunity to mobilize resources).

Vendor Selection Criteria (Draft)

Initial discussion with data provider have already revealed certain criteria which will be necessary to be considered during the selection of the right data provider:

Data Features

- Historical data and maximum number of results offered.
- Keywords search related to hate speech subject is allowed (Talk Walker for example does not allow it as per its internal compliance policy)
- The platform should have access to social media api such as Twitter Firehose and/or Twitter historical power track, Facebook pages api, etc. they should also be crawling news websites, blogs and other webpages from different sources on their servers.
- Multiple search language permitted. Arabic, English are mandatory.
- Provide data by location, and it is important to offer sub-country disaggregation even if the sample becomes too narrow, since few users enable their GP (as per discussion with a few data providers).
- Exporting the whole quantitative and qualitative data (including the tweet content) without limitation.

Data Provider functionalities

- Their system is powered by advanced search, charts and quantitative data visualization, without keywords restrictions.
- Adequate number of allowed Boolean queries on their systems including the exclude option.
- Advanced filtering criteria to filter the data noise and remove fake/robotic accounts.
- API system to link the query result to UNDP external dashboard to show findings
- AI module ready to help eliminate the data noise based on relevancy. The AI module (if possible) should allow for integrating the model developed by UNDP team.

One United Nations Plaza, New York, NY 10017,

Empowered lives.
Resilient nations.