

VOTE BALLOT



DEMOCRACY
REPORTING
INTERNATIONAL

VOTE

FEBRUARY 2023 - NOVEMBER 2023

ONLINE PUBLIC DISCOURSE IN THE MENA REGION

ANTI-IMMIGRANT HATE
SPEECH, AI SOLUTIONS,
DETECTING ONLINE
VIOLENCE AGAINST WOMEN,
AND REGIONAL STRATEGIES



Warning:

Social media
monitoring reports
contain potentially
disturbing content
that may be
distressing for some
readers.

Democracy Reporting
International is
sharing this content
only for scientific and
research purposes.

This fourth report has been produced by DRI and its partners for the project “Words Matter”. The report covers the period from February 2023 - November 2023:



DRI Partners



Jordan
Open Source
Association

Supported by



December 2023

This report is available under a public Creative Commons license. Attribution 4.0 international



Acknowledgments

We extend our heartfelt appreciation to the dedicated individuals and teams that played pivotal roles in the success of this project. The current project team, including Emna Mouelhi, Project Coordinator; Mejda Souissi, Project Officer; Wafaa Heikal, Social Media Analyst; Ikram Hajji, Data Analyst; Wafa Hmadi, Project and MEL Coordinator; Walid EL Rageiag, Finance and Administration manager; and Wael Abu Anzeh, Project Manager, who demonstrated unwavering commitment and expertise in driving the project forward.

Former colleagues, including Hervé de Baillenx, DRI MENA Representative and Project manager; Amira Kridagh, Project and MEL Coordinator; Makrem Dhifali, Data Analyst; and Mohamed Abderahim Ben Salem, Finance Coordinator, provided invaluable insights and contributions during earlier stages of the project.

Lena-Maria Böswald, from DRI's Berlin headquarters' Digital Democracy Unit, who consistently offered technical advice and feedback, greatly enriched the project's development. We extend our

appreciation to the Digital Democracy Unit at DRI's headquarters in Berlin for their continuous support.

Our heartfelt thanks also go to the DRI Tunisia office team for their significant support throughout the project, contributing to its overall success. We acknowledge the essential contributions of our finance colleagues at DRI's headquarters in Berlin – Lorand Gyenge, Finance Coordinator; and Yousef Musarsaa, Finance Coordinator, whose efforts and supervision ensured the project's financial stability.

We extend our gratitude to our esteemed project partners, whose dedication and collaboration significantly enhanced the project's impact, including Mourakiboun and the Institut de Presse et Sciences de l'Information (IPSI) from Tunisia, who comprised the "LabTrack" project; the Maharat Foundation from Lebanon; and the Al-Hayat Center – RASED for Civil Society Development and the Jordan Open-Source Association (JOSA) from Jordan. We also deeply appreciate the contributions of our partner, the Sudanese Development Initiative (SUDIA), during their

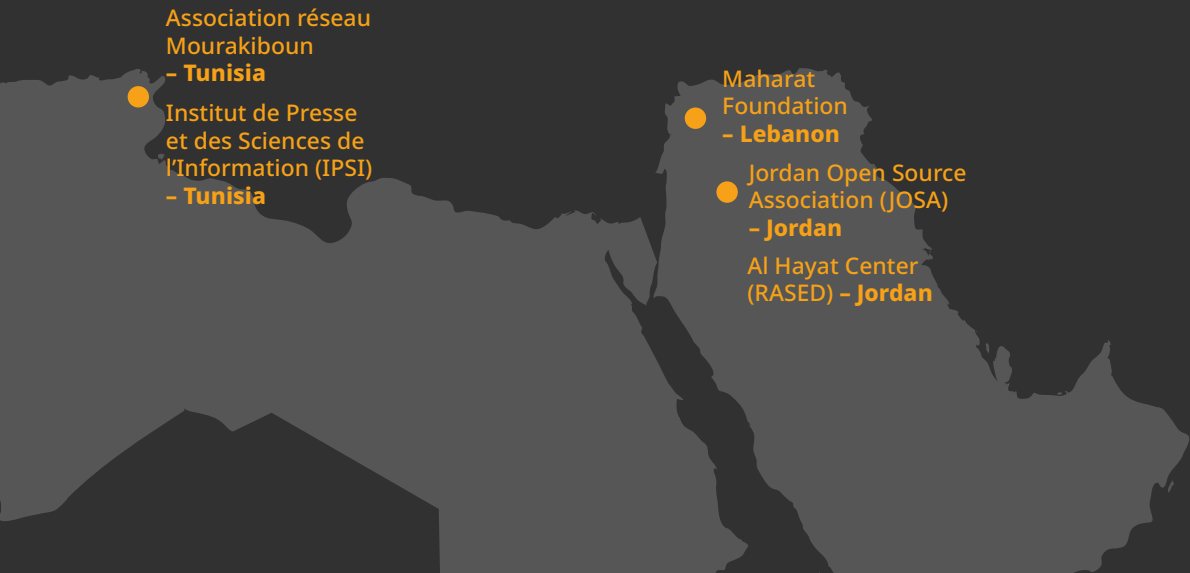
active involvement, which they had to suspend due to the beginning of the armed conflict in Sudan in April 2023. Despite the challenges posed by the ongoing conflict in Sudan, their dedication and expertise have left a lasting impact on the project's progress, and we extend our hopes for a peaceful resolution to the conflict in their country.

We are also very grateful for the support of the German Ministry of Foreign Affairs, whose funding made this project possible. Their commitment to promoting democracy and addressing critical issues in the MENA region has had a profound influence on the project's success.

We acknowledge and honor all individuals and organisations whose contributions, though not immediately apparent, were essential to the project's success. Your support, guidance, and expertise were invaluable in achieving our objectives, and your collaborative efforts have made a lasting impact.

Index

I.	Introduction	08
II.	Executive Summary	10
III.	Researching Hate Speech and Offensive Language against Sub-Saharan Immigrants in Tunisia – A Case Study	14
	1. Introduction	4
	2. Context	15
	3. Methodology	17
	4. Key Finding and Data Analysis	20
	5. Trends of hate speech and anti-immigration narratives	28
	6. Social Media Tactics Used in Spreading Anti-immigration Narratives	31
	7. Building a Lexicon of Hate Words Used in the Tunisian Dialect	33
	8. Conclusion and Recommendations	35
	9. Recommendations	35
	10. Annexes and References	37
IV.	Online Gender Violence Against Human Rights Defenders in Jordan: Hala Ahed – Case Study	44
	1. Introduction	44
	2. Context	45
	3. Methodology	46
	4. Content Analysis	50
	5. Conclusion and recommendations	57



V.	Detecting Online Gender- Based Violence: The Nuha AI Initiative	62	VI.	Words Matter Network Regional Forum Report: Expert Insights on MENA's Digital Landscape	88
	1. Introduction	62		1. Executive Summary	88
	2. Problem Statement and Objectives	62		2. Introduction	89
	3. Data Collection and Preprocessing	63		3. Day One Sessions	90
	4. Dataset Annotation	65		4. Day Two Sessions	91
	5. Model Architecture	70		5. Day Three Sessions	93
	6. Model Evaluation	72		6. Cross-Session Themes	95
	7. Limitations	72		7. Recommendations	96
	8. Ethical Considerations	76		8. Conclusion	97
	9. Future Improvements	79		9. Acknowledgments	97
	10. Conclusion	80	VI.	About Words Matter	98
	11. Recommendations and Lessons Learned	80	VII.	About the Digital Democracy programme	99
	12. References	82	VIII.	About DRI	100
	13. Glossary	84	IX.	About DRI partners	101



Introduction

The “Words Matter” project proudly presents its fourth and final report, marking a major milestone for the project. This report builds on the rigorous research, collaboration, and dedicated efforts of our partners from Jordan, Lebanon, Sudan, and Tunisia. Since its inception in February 2021, the Words Matter project has embarked on a journey aimed at safeguarding democratic processes and enhancing society’s resilience to online hate speech and disinformation in the MENA region.

The collaborative dedication of the Words Matter Network partners has been instrumental in drawing a comprehensive regional map of the digital landscape, accurately outlining trends and tactics prevalent in online manipulation, particularly during electoral periods. This collective effort has been driven by a strong commitment to combatting online violence targeting women in public digital spaces, a challenge that remains a key focus of our work.

The centrepiece of our project revolved around creating an understanding of the impact of online hate speech on societal well-being, recognising its potential to trigger psychological and physical violence. The Words Matter Network has articulated comprehensive recommendations through its

publications, stressing the importance of comprehensive solutions to create healthy information ecosystems. This approach extends beyond engagement with tech companies and reaches the grassroots level to foster resilience and combat misinformation effectively.

In close collaboration with our local partners, this report consolidates the findings from various initiatives, including case studies, AI-driven solutions, regional forums, and thorough research. Key findings highlight the pivotal role of social media in shaping public opinion, not only during elections and civil events, but also in influencing attitudes towards minorities and marginalised communities.

Individual reports from our partners delve into specific cases. Our Tunisian partner Mourakiboun in close cooperation with DRI, investigates hate speech against Sub-Saharan immigrants in Tunisia. Our partner from Jordan, JOSA examines online gender-based violence against human rights defenders in Jordan, and showcases the Nuha AI Initiative combating gender-based violence online. These case studies reflect the different forms and tactics of online hate speech, emphasising the urgent need for collective action, regulation, and societal inclusivity.

Moreover, the 2023 Words Matter Network Regional Forum brought together experts and stakeholders to dissect the MENA region's digital landscape. Discussions revealed the challenges of managing online violence, disinformation, and hate speech, calling for collaborative efforts and enhanced media literacy to create a responsible and equitable digital environment.

The collective work of the Words Matter Network, which is presented in its members' reports, in addition to the insights from the Forum, highlight the pressing need for a multi-stakeholder approach, championing human rights, fostering inclusivity, and addressing legal, technological, and educational domains. The emphasis remains on preserving freedom of expression while curbing online harm, while the commitment to research and analysis continues to pave the way for a healthier and more responsible digital landscape across the MENA region.



Executive Summary

Our partners Mourakiboun and IPSI (LabTrack), from Tunisia, in close collaboration with DRI, present a case study on hate speech against Sub-Saharan African immigrants in Tunisia. The study covers trends in hate speech from February to July 2023, particularly following a speech by Tunisian President Kais Saied and a violent incident in Sfax involving local residents and migrants. The report highlights the significant impact of social media on shaping public opinion and exacerbating discord through the spread of false information. It emphasises the historical discrimination faced by Sub-Saharan Africans in Tunisia, citing a surge in hate speech after President Saied's speech, resulting in deadly clashes in Sfax and condemnation from human rights organisations and the UN Committee on the Elimination of Racial Discrimination. The report's analysis of 27,465 Facebook posts revealed that 4.35 per cent contained hate speech. It identified narratives accusing immigrants of attempting to seize control of Tunisia, inciting violence, and involvement in theft, and criminal activities. Tactics included

manipulating Meta's guidelines and using emotionally charged language. Recommendations call for collective action, education, dialogue, legal reforms, economic initiatives, and tech measures to counter hate speech, protect human rights, and promote societal harmony. It stresses the urgent need for collaborative efforts to combat hate speech and racism against Sub-Saharan Africans in Tunisia, to foster a more inclusive and tolerant society.

A report by our partner, the Jordan Open Source Association (JOSA), examines the campaign of online gender-based violence (OGBV) against human rights defender Hala Ahed in Jordan. It highlights the hate speech and threats Ahed faced due to her advocacy for gender equality and women's rights and freedom of expression. The report contextualises Ahed's case within the context of shrinking civic space and political strife in Jordan, including government actions against the Jordanian Teachers' Syndicate. Using data annotation tools, the report delves into the hate speech campaign against Ahed, revealing

sexist, homophobic, and religiously motivated attacks against her and feminist ideals. The report identifies three primary trends within the hate speech campaign: sexist and homophobic rhetoric, demonisation of feminists and feminism, and hate speech rooted in religious beliefs. Examples of comments and reposts highlight misconceptions surrounding feminism in Jordanian society. The report urges concerted efforts to combat online hate speech and to protect fundamental rights, proposing multidimensional solutions involving stricter regulations, platform accountability, user education, victim assistance, media literacy, accountability measures, and collaborative research. It advocates for comprehensive actions and collaborations among tech corporations, government agencies, civil society, and individual users to create a safer digital environment, supporting individuals like Ahed in their pursuit of justice and equality.

In a separate report, JOSA presents its newly developed and innovative AI-based tool. The Nuha AI Initiative by JOSA is aimed at detecting and combatting OGBV against women in Jordan. Nuha, an Arabic term meaning "mind" or "brain," is designed to address the rising concern of online hate speech and violence targeting women, particularly on social media platforms. The tool seeks to enhance the safety and inclusivity of the digital space for women, by monitoring and reporting OGBV. To

develop Nuha, JOSA collaborated with the TAMAM Coalition, identifying 60 women's accounts in Jordanian public spaces for data collection. Extensive data from Facebook and Twitter, comprising 85,000 comments, underwent meticulous annotation to classify online gender-based hate speech. The report details the technical aspects of Nuha, including its architecture, training process, and evaluation metrics. Despite challenges such as data imbalance and Twitter API limitations, Nuha achieves a 72 per cent F1 score in identifying hate speech, reflecting its effectiveness. The report concludes with recommendations for digital archiving, transparency from social media platforms, increased data accessibility, industry-academia collaboration, partnerships with rights organisations, investment in AI expertise, and considerations for data sample sizes. The Nuha AI Initiative represents a significant step in addressing OGBV, showcasing JOSA's commitment to leveraging technology for social good.

Lastly, we present the main outcomes of the 2023 Words Matter Network Regional Forum, a three-day event focusing on disinformation and hate speech in the MENA region. The forum brought together experts, organisations, and stakeholders to examine the dynamics shaping the region's digital landscape, with sessions dedicated to countering online violence, promoting freedom of expression, addressing hate

speech, and analysing the information environment. Key recommendations emerged from the event, emphasising the need for legislation to combat disinformation and hate speech. Concerns were raised about the misuse of penal codes against journalists and activists, underscoring the importance of safeguarding freedom of speech in the face of legal challenges. Discussions highlighted the role of social media platforms, regulatory hurdles related to algorithms, the significance of understanding cultural contexts, and the imperative of inclusivity in digital spaces. Managing online violence, disinformation, and hate speech pose substantial challenges, necessitating collaboration and enhanced media literacy. The report outlines a spectrum of recommendations spanning legislation, collaboration, media literacy, tech accountability, and regional cooperation. These recommendations include the establishment of fact-checking protocols, the promotion of media independence, and fostering collaboration among stakeholders to effectively combat online harm. The prevalence of disinformation and hate speech emerged as critical challenges, prompting the call for tailored solutions within unique cultural and political contexts. The

report underscores the need for a comprehensive, multi-stakeholder approach that bridges legal, technological, and educational domains to address these issues while preserving freedom of expression. It highlights the complexities of addressing online harms in the MENA region, and underscores the importance of ongoing collaboration and actionable strategies to foster a responsible and equitable digital landscape.

In conclusion, the Words Matter Network findings provide insights into the challenges of disinformation, hate speech, and online violence in the MENA region. These reports highlight the urgency of collective action and comprehensive measures to combat these threats, emphasising the importance of safeguarding human rights, promoting media literacy, and fostering a more inclusive and responsible digital landscape in the region.

First Report

Researching Hate Speech and Offensive Language against Sub-Saharan Immigrants in Tunisia – A Case Study

Introduction

For this report, in collaboration with Democracy Reporting International (DRI), from February to July 2023 the "Lab Track" team monitored and analysed anti-immigration narratives on Facebook attacking sub-Saharan African migrants in Tunisia.

The first section of the report focuses on a speech by Tunisian President Kais Saied on 21 February 2023 during a meeting ¹ with the National Security Council which resulted in the dissemination of false news, hate speech, and offensive language on social media platforms.

The second section of the report pertains to the killing of a Tunisian individual in the city of Sfax during clashes between local residents and undocumented migrants from Sub-Saharan Africa, on 5 July 2023.

It is clear that social media platforms have become influential tools for shaping public opinion towards specific ideologies, and reinforcing these through false news, which can distort public perceptions, sow discord, and promote hatred among individuals and communities. Succumbing to such "news" without thoughtful consideration can lead to instilling fear, promoting

¹ (Sub-Saharan African countries repatriating citizens from Tunisia after 'shocking' statements from country's president, s.d.)

discord and hatred, and undermining mental well-being.

In this context, the project initiated by the "Mourakiboun" network, in collaboration with the Institute of Press and News Sciences in Tunisia and DRI, aims to uncover the methods and techniques employed in deceptive campaigns on Facebook pages, to analyse the type and content of the most-circulated posts, and to assess their impact on viewers.

Context

For years, sub-Saharan Africans in Tunisia have faced a range of challenges, including discrimination, racism, and hatred. According to both local and international reports,

people with black skin in Tunisia often experience harassment by security forces, arbitrary detention, and violence, along with a lack of access to healthcare, education, and job opportunities. More than 20 organisations, including the Tunisian Forum for Economic and Social Rights and the Tunisian Human Rights League, released a joint statement stating that security forces are conducting a campaign against migrants, with over 300 migrants, including women and children, being detained in centres "without adhering to proper procedures".

Social media platforms in Tunisia have witnessed an intense campaign of hate speech and offensive language

against migrants from Sub-Saharan countries since the beginning of 2023. Recently, this can be largely attributed to the statements made by the Tunisian President Kais Saïed in his statements to the National Security Council. In his speech ², the president described the alleged illegal border crossings from sub-Saharan Africa into the country as a "criminal enterprise hatched at the beginning of this century to change the demographic composition of Tunisia." Saïed said the continuous illegal immigration aims to turn Tunisia into "only an African country, with no belonging to the Arab and Muslim worlds," adding that those behind this scheme are involved in human trafficking. This statement was published by the Tunisian presidency on its [official Facebook page](#). His speech sparked significant public debate and led to an escalation of violence and hate speech against sub-Saharan African migrants in Tunisia.

In this context, there were deadly clashes between locals and migrants in the city of Sfax in July. These clashes resulted in the death of a Tunisian citizen and led to the spread of anger and incitement to violence in Sfax and other Tunisian cities.

Several human rights organisations, such as "I Watch" and "Lawyers Without Borders," as well as the Tunisian Journalists' Syndicate, condemned his speech. The UN Committee on the Elimination of Racial Discrimination ([CERD](#)) said it was alarmed by the remarks made by the President. The

² Xiaofei Xu and Kareem El Damanhoury, "Sub-Saharan African Countries Repatriating Citizens from Tunisia after 'Shocking' Statements from Country's President", CNN, 4 March 2023 – CNN

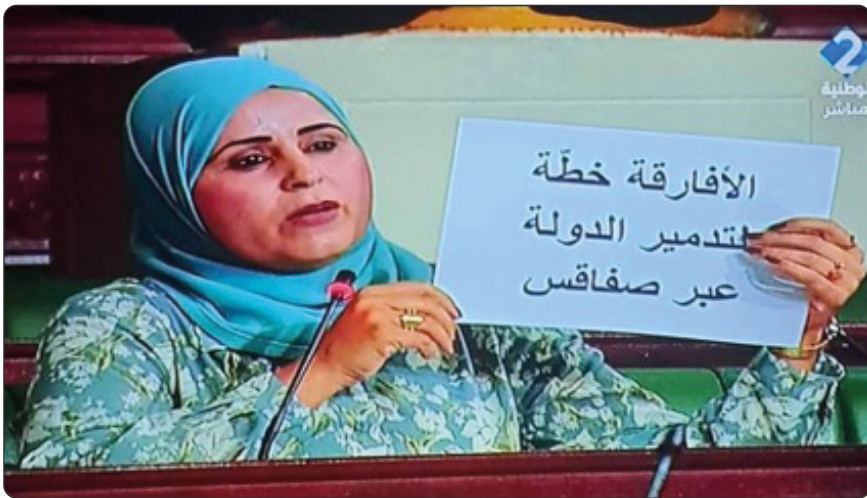
[statement](#) said that such remarks violate the International Convention on the Elimination of All Forms of Racial Discrimination.

Sfax hate crime incidents

On July 3rd, Tarak Mahdi, a member of the Tunisian Parliament, livestreamed on his [Facebook page](#)³ the aftermath of an altercation between a Tunisian man and three others, who he said were sub-Saharan migrants in the city of Sfax.

Journalists from France 24 [reported on about a dozen videos filmed and shared online on this shocking night](#), and reported that *“during the night of July 3, Tunisians attacked the homes of sub-Saharan migrants in Sfax, the country's second-largest city. The violence was sparked by the death of a Tunisian, blamed on three sub-Saharan Africans. Footage filmed by the assailants and residents shows the outbreak of violence. The police then picked up many of the migrants and abandoned them in the desert.”*

The hate crime incidents are still under investigation by the Tunisian legal system. In a further sign of the growing problem of racism against sub-Saharan Africans in Tunisia,



[Link](#): A Screenshot from the Facebook Page of Member of Parliament Manal Bdeda, 26 July 2023. The MP held a sign saying “The (migrant) Africans are a strategy to destroy the country through Sfax” during a parliament session attended by the interior minister that was broadcasted on National Tunisian National channel.

³ Tarak Mahdi [Facebook Page](#), 3 July 2023.

an MP held aloft a sign claiming migrants are part of a plan to destroy the country.

Methodology

The methodology employed in this research was data-driven, utilising a systematic approach to collecting, analysing, and classifying social media data from Facebook related to the phenomenon of anti-immigration hate speech.

Data Collection

1. Sample Selection :

The sample selection process involved the creation of a keyword list and specific timeframes, enabling data gathering from the Facebook landscape using "CrowdTangle", a social media analytics tool that allows for data collection and analysis.

For this research, the team did not specify a list of pages or criteria for selection. Instead, using advanced search capabilities, the CrowdTangle tool crawled all Facebook posts and extracted those mentioning one or more of the specified keywords within the specified time interval. This methodology helped identify the entire landscape involved in the anti-immigration hate speech campaign, without targeting specific concerned parties.

Ahead of the data collection approaches, the team identified relevant keywords encompassing racial hate speech, derogatory language, geographic references

(related to Sfax), and discriminatory terms.

These keywords served as the basis for searching and extracting relevant posts and comments on Facebook.

Additionally, the team pinpointed specific time periods to concentrate on, allowing it to analyse the evolution of hate speech trends over time and capture any noteworthy changes in the discourse.

2. Keywords used in data collection included "Africans – الأفارقة" " Ajasis – الأجاسيون (ethnic group)", and "Black – الأسود".

A full list of keywords can be found in Annex I

3. Timelines

These dates and events provided the context for the timeline and developments related to the issue of migrants from sub-Saharan African countries in Tunisia, including President Saied's statements, the response from African countries, incidents involving migrants, and societal reactions:

1. **21 February 2023:** President Qais Saied holds a meeting with the National Security Council and delivers a speech regarding migrants from sub-Saharan African countries.
2. **25 February 2023:** A boat carrying migrants sinks off the Italian coast.
3. **5 March 2023** President Saied makes statements about migrants

from sub-Saharan African countries, leading to the return of citizens from four African countries.

4. **6 March 2023:** Escalation of violence in Sfax following clashes between locals and undocumented migrants.
5. **6 March, 2023:** Tunisia announces measures ⁴ in favour of migrants from sub-Saharan African countries, which included relaxation of visa procedures and increased health and social assistance, in response to growing incitement to violence against them.
6. **12 April 2023:** The World Bank suspends dealings with Tunisia, due to attacks on African migrants following President Saied's statements.
7. **25 June 2023:** Recovery of the bodies of ten migrants from sub-Saharan Africa after their boat sinks off the coast of Tunisia.
8. **5 July 2023:** Hundreds protest in Sfax against the "growing presence" of undocumented migrants; heightened violence in Sfax after the killing of a Tunisian during clashes with migrants.

Based on this, the team chose 2 main periods:

PERIOD	TIMEFRAME	CONTEXT
1 st	February – May	President Saïd's speech and the rise of hate speech
2 nd	June – July	Clashes in Sfax and hate crimes

Data annotation

The collected data underwent annotation and categorisation based on the type of speech that each example represented, where each post and text was analysed and the category was chosen based on the context of text.

A team of annotators, possessing a profound understanding of the Tunisian context, was entrusted with this task. To ensure meticulous verification, the annotators scrutinised each text, cross-referencing the available URLs and page names linked to the posts. By examining the provided posts, the annotators were able to distinguish between expressions that can be categorised as hate speech and expressions that constitute offensive language but do not meet the criteria for hate speech based on the following classification (See Annex II).

⁴ Africanews, "[Tunisia Announces Measures to 'Improve' Life of Foreign Nationals](#)", 6 March 2023.

The guidelines were based on a study (Babak Bahador –2020) published on [the social science research council intensity scale](#) (Annex III), with examples where annotators considered it while categorising the data under six categories (See Annex II), identifying hate speech idioms, slurs, and insulting words in Arabic or English. The word “Racisme” from French is the only non-Arabic/English word that we have added to the list of words as it occurred many times and we had to include it given the importance of its indications in the analysis and weight in the words list. The word is often used as part of the Tunisian Arabic dialect as it is influenced by the French language which is considered the second language in Tunisia.

In the guidelines provided to the annotators, they were instructed to carefully review each post to determine whether it related to anti-immigrant hate speech. If a post did relate to such speech, it was to be classified as “Neutral” for later analysis.

When a comment contained hate speech or offensive language, it was marked as “Yes.” In such cases, annotators were advised to evaluate the entire text, considering sentence context and incorporating the local Tunisian context to understand offensive or hate speech language in Arabic or English. The annotators were then tasked with categorising

the text under one of six categories (hate speech/offensive language), identifying hate speech idioms, slurs, and insulting words, translating hate speech words from Tunisian dialects into Arabic or English, distinguishing between offensive and hate language, and noting that some posts might fit multiple categorisations, which should be accounted for in the data analysis.

At the end of the data-annotation step, a lexicon comprising the most frequently used anti-immigration hate speech terms was compiled, which proved valuable during the subsequent data analysis phase.

The tool employed for data annotation was Label Studio, an open-source and freely available machine learning tool designed for categorising data into distinct categories.

Data analysis

The approach employed in this report is content analysis, aimed at contextualising hate speech occurrences to gain a deeper understanding of the underlying issues and motivations driving such discourse.

The primary objective of this analysis is to comprehend the narratives prevalent within anti-immigration hate speech campaigns, and to construct a lexicon to facilitate future reports or projects. It is worth noting that all data analyses will be focused on the text content of the posts.

As part of our analytical toolkit, Python algorithms were utilised for text pre-processing, encompassing tasks such as data cleaning, occurrence extraction, and similarity identification. Furthermore, Power BI was employed for generating visualisations to enhance the presentation of our findings.

Key findings and Data Analysis

Key findings

General key findings

During the observation period, which lasted from February 21 to July 15, 2023:

- **27,465 posts were monitored on 2602 Facebook pages**, including those of media outlets and various random pages related to the crisis surrounding Sub-Saharan Africans in Tunisia.

MEDIA_NEWS_COMPANY	216	NGO	47
ACTIVITY_GENERAL	210	TOPIC_NEWSPAPER	42
NEWS_SITE	206	POLITICAL_PARTY	40
COMMUNITY	161	NON_PROFIT	38
PERSONAL_BLOG	132	SPORTS	35
PERSON	88	ARTIST	34
POLITICIAN	88	TV_CHANNEL	32
JOURNALIST	80	DIGITAL_CREATOR	30
RADIO_STATION	73	MEDIA	29
LOCAL	63	TOPIC_JUST_FOR_FUN	27
BLOGGER	60	SPORTS_TEAM	26
MAGAZINE	60		
GOVERNMENT_ORGANIZATION	50	Total	2602

Table 1: Categories of pages monitored in the data-collection phase

- **Of the 27,465 posts, 1,195 posts (4.35 per cent) were categorised as containing hate speech, offensive language, or both:**
 - 2.96 per cent were classified as hate speech.
 - 1.03 per cent were categorised as offensive language.
 - 0.0146 per cent fell into both the hate speech and offensive language categories.
 - In addition, 0.21 per cent of the posts were labelled as "fake", as they contained misinformation, according to the annotators.

Neutral	26306	95,78%
Hate speech	814	2,96%
Offensive language	283	1,03%
FAKE	58	0,21%
Offensive language, Hate speech	4	0,01%
Total	27465	100,00%

Table 2: Classification of content based on the categories identified in the data-annotation guidelines

- **42.01 per cent of the content appeared in the form of photos, while 35.37 per cent consisted of links to videos or articles.** Text-based posts make up only 5.14 per cent of the total, which posed a challenge in identifying and detecting hate speech content.

Photo	11538	42,01%
Link	9715	35,37%
Native Video	2977	10,84%
Status	1412	5,14%
Live Video Complete	1407	5,12%
Youtube	334	1,22%
Video	66	0,24%
Live Video Scheduled	16	0,06%
Total	27465	100,00%

Table 3: Type of content monitored

Hate speech and offensive language key findings

During the observation period, which lasted from February 21 to July 15, 2023:

- **27,465 posts were monitored on 2602 Facebook pages**, including those of media outlets and various random pages related to the crisis surrounding Sub-Saharan Africans in Tunisia.
- **70,23 per cent of the 4.35 per cent of examples** containing hate speech or offensive language, or both, was identified as hate speech content
- **45.38 per cent of the content identified as containing hate speech or offensive language was in the form of photos, while 22.52 per cent consisted of videos.** Text-based posts make up only 13.37 per cent of the total, which posed a challenge in detecting the narratives and the language used in the hate speech campaign.
- **40.64 per cent** of the pages involved in the hate speech and offensive language campaign against Sub-Saharan immigrants are pages that are not run by organisations or official bodies, and were labelled with the following Facebook categories: **“Activity General, Personal Blog, Topic just for fun, Comedian, Person”**

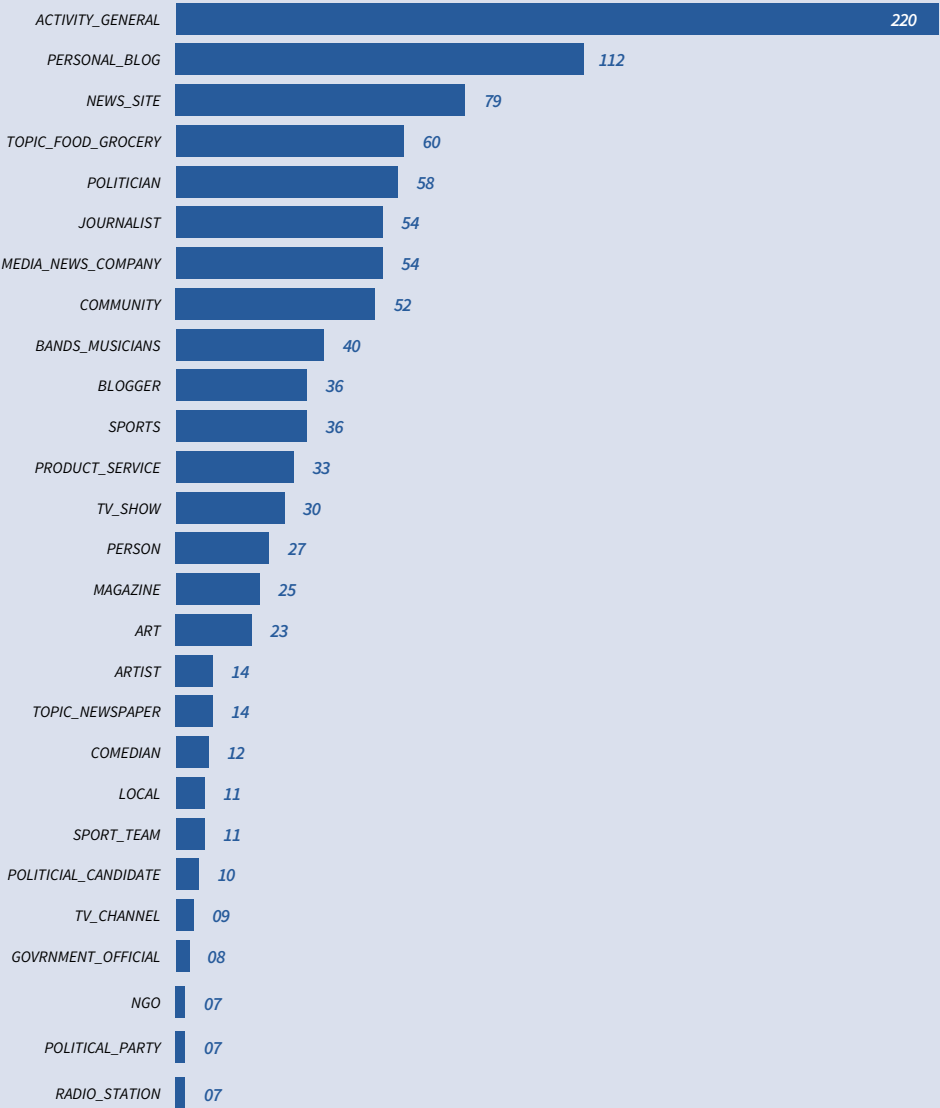


Figure 1: Page categories involved in the anti-immigration campaigns

- **90 out of 1,159 posts for Facebook pages checked on 22 February** contained hate speech and offensive language were detected. This day saw the highest number of published posts and the highest number of interactions (likes, comments, shares...). It is noteworthy that this date right after President Saïed's speech in our timeline.
- **61.97 per cent (719) of hate speech & offensive language content** was detected in the **first period** identified in the data collection section in the methodology (February - May).
- 38.48% of all hate speech and offensive language content in posts in both date periods was identified in February (See Table 4).

A potential reason for the concentration of comments in this period is that it was in the immediate aftermath of the statements by President Saïed, when anti-immigrant sentiment was at its highest. But later, with the appearance of the economic repercussions of the president's statements, when the World Bank cut off loans, and with the news of the discovery of bodies of dead migrants, the government reduced its inflammatory statements, and anti-immigrant sentiments were reduced.

MONTH	NUMBER OF POSTS	PERCENTAGE
February	446	38.48%
July	293	25.28%
June	147	12.68%
March	114	9.84%
April	96	8.28%
May	63	5.44%
Total	1,159	100%

Table 4: Number of posts containing hate speech or offensive language per month

- Out of 1,159 posts** containing hate speech and offensive language, **65.28 per cent (757)** were about Tunisia as a general demographic place. Additionally, **26.64 per cent** of these posts specifically mentioned Sfax, which was the primary focus in the second period. The identification of governorates in the textual content was based on two languages (Arabic and French), as the posts and the Tunisian dialect in general are a combination of these languages. The calculation involved summing the relevant data.

GOVERNORATES	ENGLISH	OCCURRENCES
تونس	Tunisia	757
صفاقس	Sfax	309
المهدية	Mahdia	16
مدنين	Medenine	14
سوسة	Sousse	8
المنستير	Monastir	8
سيدي بوزيد	Sidi Bouzid	6
القصرين	Kasserine	6
نابل	Nabeul	5
جندوبة	Jendouba	5
منوبة	Manouba	3
بنزرت	Bizerte	2
باجة	Beja	2
الكاف	Kef	2
سليانة	Seliana	2
القيروان	Kairouan	1
قابس	Gabes	1

Table 5: Number of posts containing hate speech or offensive language, per governorate

Data analysis

Hate speech and offensive language directed towards Sub-Saharan Africans accounted for approximately 5 per cent of the content within the dataset. At first glance, this percentage may appear relatively low, but several key factors contributed to this finding:

- Facebook employs robust content moderation systems and community guidelines to address hate speech. These systems often result in the timely removal of offensive content, either by the platform itself or by page administrators. As a result, hate speech is not always highly visible or persistent in the public domain. Facebook isn't consistent in applying such moderation systems, and it depends on Meta's resources in the region.
- Another significant challenge in gauging the true extent of hate speech pertains to private groups and comments on posts. Such content is often concealed within these private spaces, making monitoring and assessment considerably more difficult. Private groups typically have strict privacy settings and limited accessibility, which renders their content less visible to external observers. Additionally, comments on posts can also contain hate speech, but tracking and assessing such content can be challenging, due to its scattered and context-dependent nature. Regrettably, our report lacks comprehensive data on hate speech within private groups and comments on posts, primarily because of these inherent limitations.
- In our data annotation process, it is crucial to note that our primary focus was on identifying direct hate speech that utilises strong, offensive language. Not all criticism of or commentary on immigration policies per se, however, falls under the umbrella of offensive language or hate speech. Constructive criticism, policy discussions, and informed commentary on immigration policies are vital aspects of public discourse, and they should not be confused with offensive or hateful content. Our team made a clear distinction between such legitimate discussions and more extreme and offensive expressions, ensuring that our analysis accurately reflects the presence of direct hate speech, while allowing for open and constructive dialogue surrounding immigration policies. This delineation is essential for providing a balanced and accurate representation of the data.
- As mentioned in the key findings, most of the content consists of photos. It can be difficult to detect whether a post contains hate speech, offensive language, or is neutral, especially when there is no accompanying text. This is one of the limitations of the data annotations.

Although the percentage of detected anti-immigration hate speech and offensive language was low, the team focused on conducting contextual analysis of the percentage of hate speech and offensive language. This analysis involved the detection of narratives and trends, and the creation of a lexicon of hate speech.

In the data annotation, the team tried to label posts that were considered as hate speech and offensive language into the subcategories identified earlier in the guidelines, based on the social science research council intensity scale mentioned in the methodology section, and these sub-labels helped in the narrative analysis

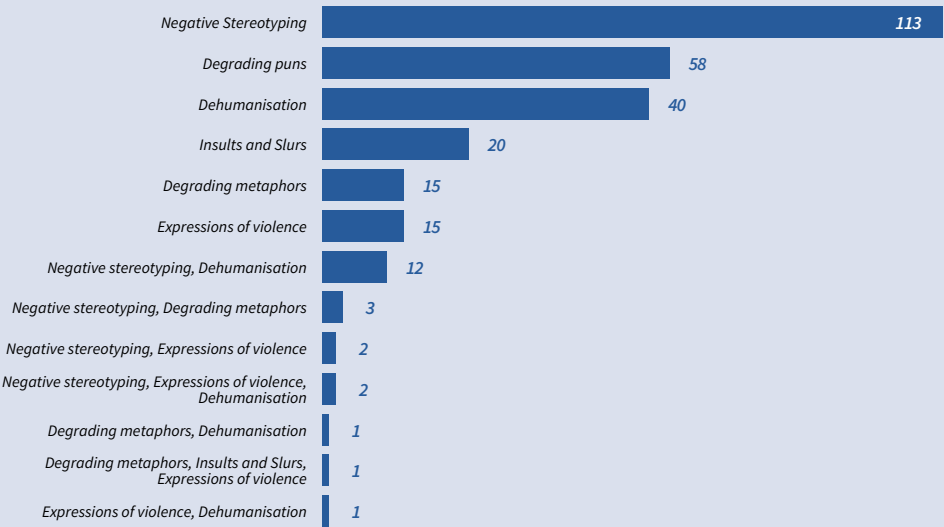


Figure 2: Subcategories and number of posts for each subcategory

In the main periods monitored, there was a surge in hate speech campaigns in Tunisia against sub-Saharan immigrants. These campaigns have been manifested in various narratives, with some posts claiming that immigrants from sub-Saharan Africa aim to take over Tunisia, engage in violent acts and attacks, or are involved in theft and damage to stores. Additionally, another narrative focuses on the discrimination faced by those with darker skin tones, even if they are Tunisian citizens residing in regions like Medenin, Gabes, and the south of Tunisia, where they are subjected to bullying and prejudice, due to their skin colour.

Trends of hate speech and anti-immigration narratives

Accusations of Immigrants' Intentions

Posts circulating in the Tunisian social media landscape suggest that sub-Saharan immigrants harbour intentions to claim Tunisia as their own.



Screenshot 02: Screenshots of posts

1. The post claims that there are over 700,000 sub-Saharan immigrants in Tunisia, and that, with the support of European funds, they are expanding in the country, with the intention of making it their own.
2. The post claims that the critical situation resulting from the increasing number of African illegal immigrants in Tunisia is being compared to the early stages of the Palestinian conflict. There is a prevailing fear of losing control over their own country, Tunisia, and of being expelled from their homeland in the future. The

presence of these immigrants is seen as a threat to national security and to the Arabic Islamic identity. Disturbing incidents, such as the abduction of a Tunisian man and the murder of his wife, as well as numerous robberies, have already been reported.

Blaming Immigrants for Violence and Crime

Portraying sub-Saharan immigrants as responsible for violent acts and crimes in Tunisia, along with mentioning the involvement of mafia and gangs, represents a deeply harmful aspect of hate speech campaigns on social media. These false generalisations fuel prejudices, foster division, and create an unsafe environment for immigrants, as they are unfairly blamed for crimes they may not have committed. The references to mafia and gangs sensationalise the issue, stigmatise the immigrant community, and often lack concrete evidence, contributing to the misleading and divisive nature of these narratives. Addressing and countering such rhetoric is vital to promoting a more inclusive and harmonious society.

Inflammatory content also accuses sub-Saharan immigrants of theft and damaging local businesses. This not only perpetuates stereotypes, but is also used to justify hostility towards the immigrant population.



Screenshot 03: Screenshots of posts

1. A video referring to the knifing murder of a Tunisian citizen in Sfax by a group of sub-Saharan immigrants.
2. Images of destroyed cars, with text asserting that illegal immigrants are responsible for incidents such as rapes, thefts, the deaths of Tunisians, drug-related activities, and prostitution, and further claiming that the Tunisian police are unable to control these situations.

Dehumanising Discourse: Stripping Immigrants of Their Humanity

In the hate speech content found in the posts, an alarming narrative emerges, where some Tunisians not only falsely blame sub-Saharan immigrants for crimes and violence, but also dehumanise them to an extent that they are denied even the basic recognition of their humanity. This degrading dehumanisation is a deeply disturbing aspect, as it reduces sub-Saharan immigrants to mere objects of scorn and prejudice, stripping them of their fundamental dignity and rights. This extreme form of dehumanisation perpetuates a dangerous “us-versus-them” mentality and reinforces the hostile atmosphere, making it even more imperative to counter such narratives and promote tolerance, empathy, and respect for the inherent humanity of all individuals, regardless of their background.



Screenshot 04: Screenshots of posts

1. The post shows that the number of tuberculosis patients in Sfax has increased, and attributes this to Sub-Saharans arriving with diseases.
2. The post asserts that the “black immigrants” are making Tunisia trash with their corruption and “sorcery activities”.

Social Media Tactics Used in Spreading Anti-immigration Narratives

- One of the main tactics used by groups who publish anti-immigration posts is a violation of Meta's community guidelines, in the impersonation of other people. The data analysis of these pages, categorised by names and content, reveals that most of these pages use the names of Tunisian celebrities. This aligns with a trend previously observed in Tunisian social media reports by DRI, where fake pages impersonate celebrities with significant followings to spread hate speech content. This is also a consistent tactic that DRI and LabTrack documented before and during the Tunisian parliamentary elections in 2022-2023



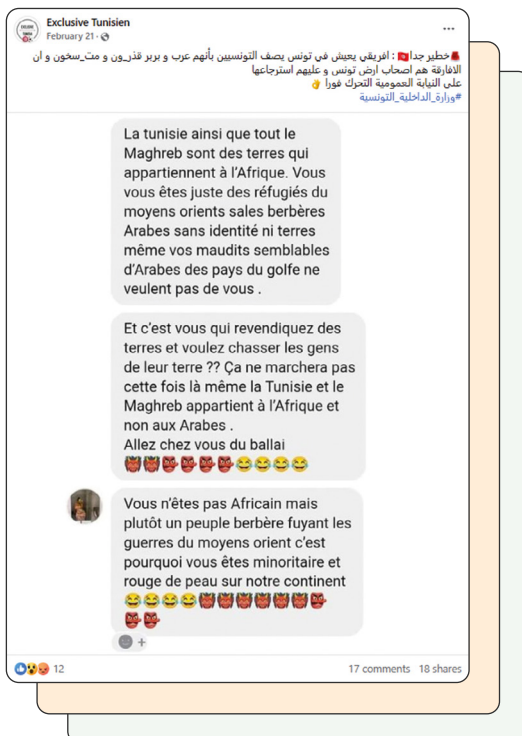
Screenshot 05: Screenshots of posts

Pages with names of a famous Tunisian actress (1) and a famous Tunisian singer (2) sharing statements against immigrants.

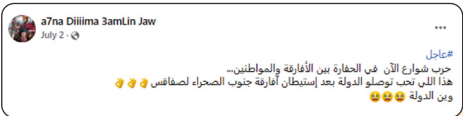
A page featuring a well-known Tunisian TV series (3) shares A page featuring a well-known Tunisian TV series (3) shares a link for a news website with the title "Urgent: African Migrants Take The Train Headed Towards Tunis"



- **One of the tactics used is posting screenshots of private conversations from private messaging platforms to public pages.** For example, one post displays screenshots of a conversation involving an African immigrant residing in Tunisia referring to Tunisians as "filthy" and "Berber", and claiming Tunisia as their homeland, which they need to reclaim. Notably, these screenshots do not reveal the names or images of the sender. Screenshots of conversations are hard to verify and debunk by fact-checkers and journalists, which makes it hard to counter such narratives.
- Most of the content in the data sample was visual. Although our resources were limited to textual, it was very clear that visual posts like viral videos and memes comprised a large percentage of the content.



- The textual analysis reveals a notable tactic, characterised by the frequent use of phrases like 'خطير' "dangerous", "danger", and "be careful" at the start of social media posts. This deliberate strategy is aimed at immediately instilling a sense of urgency and fear in the reader. While effective in capturing the audience's attention, this emotional manipulation tactic often comes at the expense of critical thinking. Furthermore, it has the potential to contribute to confirmation bias, to deepen the polarisation of opinions, and to facilitate the spread of misinformation.
- Moreover, this strategy can create a hostile environment for targeted groups, with the potential to incite real-world harm.



Screenshot 06: Screenshots of posts with the caption: "Breaking: The destination of sub-Saharan African migrants have been identified"

1. "Breaking: A street conflict erupts between African immigrants and Tunisian citizens. Is this the future we envision after Africans claim Tunisia as their land?"
2. "A cry of terror: Here are pictures of children suffering from lice and scabies and we are to blame"

Building a Lexicon of Hate Words Used in the Tunisian Dialect

In the data annotation process, after thoroughly reviewing the content of the posts and assigning appropriate labels, annotators also engaged in the important task of identifying hate speech idioms, slurs, and insulting words present within the text. This additional step is aimed at curating a "lexicon of hate" – essentially compiling a list of specific words, phrases, and expressions that are commonly used to convey derogatory, offensive, or prejudiced meanings. This hate speech lexicon serves as a valuable resource for understanding the language and rhetoric employed in hate speech content, facilitating more effective monitoring, analysis, and mitigation efforts in the fight against online hate speech.

Note that each word can appear in more than one post, which implies that these words and expressions are not isolated occurrences, but are used consistently across multiple posts or discussions. This can reinforce the presence of hate speech or offensive language in those contexts.

Tunisia abolished slavery in 1846, and it was also one of the first countries in the region to criminalise racism, by passing [Law 50](#), the “Elimination of All Forms of Racial Discrimination” Act in 2018. During research conducted to build the lexicon, however, we found that more than 50 per cent of the words counted in the lexicon refer directly to skin tone, in the form of degrading comments about black immigrants. The word “Slave, **عبد**” was repeated 157 times in the course of building this lexicon (see the full lexicon in Annex IV).



WordCloud based on occurrences in Annex III

The creation of a Tunisian anti-immigration lexicon serves the purpose of improving the algorithmic detection of harmful content in local Arabic dialects, addressing a significant challenge for algorithm systems, and especially AI algorithms used to enhance content moderation. This lexicon can play a pivotal role in enabling better recognition and understanding of problematic content specific to the region.

Moreover, there is a technical opportunity to connect this lexicon with the artificial intelligence (AI) model developed by JOSA for detecting online gender-based violence (OGBV) in the regional context. By integrating the Tunisian lexicon with JOSA's Nuha AI model, it can significantly enhance capabilities to identify and address issues related to OGBV within the Tunisian and, possibly, the broader North African context. This connection would foster a more comprehensive

and culturally sensitive approach to detecting harmful content and addressing sensitive social issues, particularly OGBV, using advanced AI technologies.

Conclusion and Recommendations

The crisis centred around Sub-Saharan Africans in Tunisia underscores the complex dynamics of modern society and the power of social media. The prevalence of hateful and offensive comments targeted at black migrants in Tunis resulted in violations of immigrants' rights in the city, taking many shapes, including the arbitrary loss of jobs or housing, an increase in cases of arbitrary detention, and testimonies of migrants threatened with physical violence, which can sometimes escalate into broader violence and killings, as witnessed in Sfax.

Although this report focused on analysing Facebook posts containing hate speech, we recognise the role of independent media in Tunis and factcheckers who curbed the misdisinformation from spreading further on platforms like Facebook.

The findings of this report shed light on the gravity of the issue. Hate speech and incitement to violence against sub-Saharan Africans not only violate fundamental human rights, but also undermine the very fabric of a diverse and inclusive society.

To address this issue, a collaborative effort is imperative. Tunisian authorities, civil society and media

organisations, and international bodies must work together to counter hate speech and promote the provision of accurate information. Public awareness campaigns emphasising tolerance, empathy, and respect for diversity are vital to fostering a more harmonious and equitable environment.

Ultimately, the way forward requires a united commitment to combating hate speech, dispelling dis/misinformation, and nurturing a society that values the dignity and rights of every individual, regardless of their background. Only through such concerted efforts can Tunisia aspire to a more inclusive and tolerant future.

Recommendations

Considering this as a serious issue requiring serious intervention, the following recommendations can help create a more tolerant and respectful environment for all members of society, emphasising the importance of promoting human values and human rights in Tunisia and beyond.

Civil Society & Media

Verification of Information: Ensure the accuracy of information and news before publishing or sharing on social media platforms. Rely on reliable sources and credible news websites to verify news and information.

Awareness and Education: Efforts should be directed towards raising awareness among all individuals about the importance of peaceful coexistence and mutual respect between cultures and races. Increasing awareness about the necessity of

maintaining social harmony can foster understanding and cooperation among different segments of society.

Research and Analysis: Encourage Tunisian researchers and provide them with opportunities to explore the phenomenon of the spread of hate and anti-immigration racist narratives using memes and videos through platforms like TikTok and Instagram.

Dialogue and Openness: Promote dialogue and cooperation between Sub-Saharan Africans and Tunisians as the primary means to resolve this crisis. Civil society, media, and other stakeholders, such as the government, should work together to create a safe and conducive environment for peaceful coexistence and mutual respect.

Government

Protection from Violence and Hatred: The Tunisian government should take responsibility for protecting all citizens and migrants from violence and hatred. Necessary protection should be provided to Sub-Saharan Africans, ensuring their rights and dignity.

Enhancement of Legal Measures: Tackle the anti-black structural racism that affects society, and carry out reforms to respect human rights and end racial discrimination.

Eradication of Poverty and Unemployment: Addressing poverty and unemployment should be a priority for the Tunisian government. Improving economic conditions and providing job opportunities can enhance the overall quality of life and reduce social tensions.

International Cooperation: Strengthen cooperation and knowledge exchange

with local and international entities and non-governmental organisations to combat hate speech and incitement to violence.

Tech and Social Media Platforms

Implement measures on social media platforms to curtail the spread of hate speech and violence, such as:

- Promoting content that encourages peaceful coexistence and intercultural harmony, while minimising offensive and extremist content;
- Raising awareness among users about the dangers of racist and hateful speech, encouraging the reporting of such content;
- Providing support for specialised research and analysis focused on monitoring and addressing the phenomenon of hate speech and violence. This includes using technology and AI to identify false news, fake accounts, and networks promoting discrimination and hatred.
- Enhancing content moderation systems, by focusing more on the lexicons of local Arabic dialects, hiring more local content moderators for each country in the region;
- Collaborating with governments, non-governmental organisations, and civil society globally to combat violence and hatred; and
- Encouraging platforms to collaborate with researchers and experts to develop effective policies and technologies to counteract racist and hateful speech.

Annexes and References







Annex 1: Keywords used in data collection for the sample

Arabic Term	English Translation
الافارقة	Africans
الأجاصيون	Ajasis (ethnic group)
ترحيل	Deportation
كحلوش	Black
كحالش	Black
الافارقة بصفاقس	Africans in Sfax
الوصفان	Blacks
هجرة الافارقة	African migration
توطين الأفارقة	African resettlement
جنوب الصحراء	Sub-Saharan Africa
السود	Black people
توافد الأفارقة	African influx
الإستيطان	Colonisation
حملات عنصرية	Racist campaigns
اعتداءات أفارقة	Attacks on Africans
اللاجئين الافارقة	African refugees
المهاجرين الافارقة	African migrants
اللاجئون الافارقة	African asylum seekers
Racisme (French word used by many Tunisians as part of the Tunisian Arabic dialect)	Racism

Annex 2: Classification of Hate Speech and Offensive Language

HATE SPEECH	
CATEGORY	DEFINITIONS
Negative stereotyping	<p>The attribution of negatively connoted characteristics, roles, or behaviours to a whole group or to individuals on the basis of their group membership.</p> <p>Verbal attacks (including harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal).</p>
Dehumanisation	<p>Statements that equate or compare humans to inanimate things, animals, or non-human beings, or characterise humans as savage or animalistic.</p>
Expressions of violence	<p>Statements that justify, incite, or threaten physical violence against or the killing of an individual or a group. Calls for exclusion or segregation made against groups can be included.</p>
OFFENSIVE LANGUAGE	
CATEGORY	DEFINITIONS
Insults and slurs	<p>A gross indignity – an instance of insolent or contemptuous speech or conduct (Merriam Webster)</p>
Degrading metaphors	<p>Metaphors that contain degrading or humiliating language. The meaning of degrading is causing or associated with a low, destitute, or demoralised state – causing someone to be or feel degraded. (Merriam Webster)</p> <p>For example: Statements that put immigrants in lower positions, and describing them with degrading examples</p> <p>مثال: يقعدوا ديما عبید، نشتریهم بالفلوس</p>
Degrading puns	<p>Using language of degradation (degrading is causing or associating with a low, destitute, or demoralised state – causing someone to be or feel degraded) in sarcastic speech.</p> <p>For example: The whole text is a pun, but it contains degrading words.</p>

Annex 3: Classifying and Identifying the Intensity of Hate Speech

Color	Title	Description	Examples
	1 Disagreement	Rhetoric includes disagreeing at the idea/belief level. Responses include challenging claims, ideas, beliefs, or trying to change their view.	False, incorrect, wrong, challenge, persuade, change minds
	2 Negative Actions	Rhetoric includes negative nonviolent actions associated with the group. Responses include nonviolent actions including metaphors.	Threatened, stole, outrageous act, poor treatment, alienate
	3 Negative Character	Rhetoric includes nonviolent characterizations and insults. There are no responses for #3.	Stupid, thief, aggressor, fake, crazy
	4 Demonizing and Dehumanizing	Rhetoric includes subhuman and superhuman characteristics. There are no responses for #4.	Rat, monkey, Nazi, demon, cancer, monster
	5 Violence	Rhetoric includes infliction of physical harm or metaphoric/ aspirational physical harm or death. Responses include calls for literal violence or metaphoric/aspirational physical harm or death.	Punched, raped, starved, torturing, mugging
	6 Death	Rhetoric includes literal killing by group. Responses include the literal death/elimination of a group.	Killed, annihilate, destroy

Annex 4: Arabic -English lexicon of Hate Words Against Sub-Sahara Immigrants in Tunisia

ARABIC WORD	ENGLISH TRANSLATION	ARABIC EXPLANATION	ENGLISH EXPLANATION	OCCURRENCES
عبيد	Slaves	بمعنى العبد الذي يتبع سيّدا	Refers to a slave who follows a master	157
مجرّمة	Criminals	الذين يقومون بأعمال إجرامية	Those who engage in criminal acts	96
همج	Barbarians	من الهمجية و مفتعلي الفوضى	Those who commit barbarism and create chaos	53
وصفان	Despicable	يُقصد بها أسمر البشرة وتوحي بمعنى الوصيف أو صاحب المرتبة للتقليل من الشأن	Referring to dark-skinned person, with the implication of inferiority	44
خايين	Ugly	قبيحين	Ugly	27
اجصيين	Inferior	مختصر لـ "أفرقة جنوب الصحراء"	An abbreviation for "Africans south of the Sahara"	25
جورة	Oppressors	قليل الشأن، الهمجي	Lowly, barbaric	19
فاسدين	Corrupt	من الفساد	From corruption	19
محتلين	Occupiers	المحتل، أو المستعمر لأرض ليست على ملكه	Occupiers or colonisers of a land they do not own	17
وصيف	Black	يقصد بيها أسمر البشرة و توحي بمعنى الوصيف أو صاحب المرتبة للتقليل من الشأن	Referring to a dark-skinned person, with the implication of inferiority	16
كحالش	Black	أسمر البشرة	Dark-skinned	15
سراق	Thieves	السارق	Thieves	12
متآمرين	Conspirators	متآمر أو من يكيد الدسائس	A conspirator or someone who schemes, conspires	11
كلوش	Filthy	الشخص الوسخ	A filthy person	10
حنّالة	Trash	حنّالة القوم	People's trash	9
إرهابيين	Terrorists	بمعنى التهريب، -من الإرهاب-	In the sense of terror, from terrorism	8
باندية	Gangsters	عصابات	To be a member of a gang to commit crimes	8
جراثيم	Germes	جراثيم (جمع جرثومة)	Germes	5

رهبانة	Scoundrels	من يتظاهر بطبع جيّد غير طبعه للوصول الى مبتغى	Someone who pretends to have a good nature, different from their nature, to achieve a certain goal	3
الطحانة	Pimps	القوَّاد	Pimps	3
كلوشارات	Filth	الشخص الهمجي أو العصابات	Tramps	3
جرايح	Vultures	توصيف بمعنى الفأر أو الجربوع	A description meaning a rat or vulture	2
خراوات	Piece of shit	قاذورات بشرية	(Human filth (piece of shit	2
الخماج	Scum	من الوسخ و التلوث	From filth and pollution	2
ناتنين	Stinking	صاحب الرائحة النتنة و الكريهة	One with a foul, unpleasant smell	1
اولاد الحفيانة	Homeless	المتشرّدون	Homeless people	1
بهايم	Beasts	التوصيف بالحيوان، قليل الذكاء	Description as an animal, of low intelligence	1
قطاع طرق	Road Blockers	قطاع طرق	Road blockers	1
حيوانيز	Animalism	مشتق من كلمة "حيوان"	Derived from the word "animal"	1
مذلولين	Humiliated	قليلو الشأن ، بمعنى ذليل ، من الذل	People of little significance	1
مئيكين	Asshole	سيء السمعة ، تستعمل للتوصيف بالانحطاط و قلة الشأن	Used to describe someone as disreputable and of little significance	1
مستخين	Dirty	الشخص الوسخ	A dirty person	1
هز طيش الكازي	Throwing	الالقاء بالشيء، تستعمل للتقليل من القيمة	Throwing something away, used to diminish its value	1
قعار	Cowards	قليل الشأن، الهمجي	Inelegant, tasteless	1
كعالف	Filth	توصيف الشخص الوسخ	Description of a dirty person	1
متع البراكجات	Road blockers	قطاع طرق	Road blockers	1
قاتلين لارواح	Killers of Souls	القاتل أو المجرم	A killer or criminal	1
وكالين القطاطس	Eaters of cat meat	أكلي لحوم القطط	Eaters of cat meat	1

References

1. (Sub-Saharan African countries repatriating citizens from Tunisia after 'shocking' statements from country's president, s.d.)
2. Xiaofei Xu and Kareem El Damanhoury, **"Sub-Saharan African Countries Repatriating Citizens from Tunisia after 'Shocking' Statements from Country's President"**, CNN, 4 March 2023 – CNN
3. **Tarak Mahdi** Facebook Page, 3 July 2023.
4. Africanews, **"Tunisia Announces Measures to 'Improve' Life of Foreign Nationals"**, 6 March 2023.
5. Paasch-Colberg, Sünje, (2022), **"Insults, Criminalisation, and Calls for Violence: Forms of Hate Speech and Offensive Language in German User Comments on Immigration."**, https://doi.org/10.1007/978-3-030-92103-3_6.
6. Ugarte, Rodrigo. **"Classifying and Identifying the Intensity of Hate Speech."** Items, items.ssrc.org/disinformation-democracy-and-conflict-prevention/classifying-and-identifying-the-intensity-of-hate-speech/.
7. **(france24.com)** Footage shows sub-Saharan African migrants being attacked and expelled over 48 hours in Tunisia <https://observers.france24.com/en/africa/20230706-tunisia-sfax-migrant-sub-saharan-raids-violence-images>

Second Report

Online Gender Violence Against Human Rights Defenders in Jordan: Hala Ahed – Case Study

Introduction

Hala Ahed is a prominent Jordanian human rights defender and lawyer, specialising in gender and women's rights, labor rights, trade unions, and freedom of opinion and expression. She has been a member of the legal team for many local organisations and groups working on human and civil rights, and has gained recognition as a member of the legal team that defended the Jordanian Teachers' Syndicate.¹ In May 2023, Ahed received the Front

Line Defenders Award, highlighting her dedication and activism.²

Ahed has faced various forms of violence and threats. In January 2022, an investigation by Access Now and Front Line Defenders revealed that her phone had been infected with NSO's Pegasus Spyware since March 2021.³ Moreover, in June 2023, she was targeted by a hate speech campaign, following a public invitation to a session on feminism and the basis of feminism, led by Ahed. After the Ahel organisation published the public invitation for the session, a wave

¹ Front Line Defenders, [Hala Ahed Deeb](#).

² Jordan News, "[Jordanian Lawyer Hala Al-Ahed Wins Front Line Defenders Award](#)", 27 May 2023.

³ Marwa Fatafta, "[Unsafe Anywhere: Women Human Rights Defenders Speak Out about Pegasus Attacks](#)", Access Now, 17 January 2022.

of online attacks started emerging, with much focus on the term “feminism”, which carries a negative connotation in Jordanian society, in the post. Examples of these can be seen in screenshots 2, 3, and 4. This negative misconception about feminism mainly comes from the idea that feminism seeks to destroy traditional family values in Jordan, and replace them with Western values. The attacks later turned into a hate speech campaign that targeted Ahed specifically.⁴

Context

Earlier this year, Freedom House, a U.S.-based NGO that monitors political and civil freedoms around the globe, categorised ⁵ Jordan as “Not Free”, with an overall score of 33/100. In recent years, the civic space in Jordan has been steadily shrinking, as a result of stricter legislation and authoritative measures. In July 2020, the government shut down ⁶ the “Jordanian Teachers’ Syndicate” (JTS) and arrested its leaders, based on allegations made on social media of corruption and “inflammatory measures”. Since its establishment, the Syndicate had been at odds with the government over many issues related to teachers’ rights and compensation. The situation reached a boiling point in September 2019, when JTS announced an open strike until its demands were met. The strike ended one month later with a deal between the syndicate and

the government, but started a series of actions by the government that were designed to diminish the role and influence of the syndicate, which culminated in it being shut down less than a year later. For the period between the strike in 2019 and the shut-down in 2020, the syndicate was headed by Naser Al-Nawasrah, a member of the Jordanian Muslim Brotherhood, which is the main political rival of the government in Jordan, and who took on this role after the originally elected president died in a car accident. This gave the struggle between the government and the syndicate a political dimension that was seen as threatening, and state media openly accused ⁷ the Muslim Brotherhood of using this strike as a mean to achieve political goals, an accusation that was repeatedly denied by the Brotherhood. The strike also created divisions between people who supported the Brotherhood and those who supported the government.

When the trials for the Syndicate’s leaders opened, a group of lawyers came together to represent defendants, among whom was Ahed.⁸ She was one of the most prominent members of the defence team, and had faced much harassment ⁹ from the Jordanian authorities due to her activities, making her a target of the security apparatus. She was also one of four human rights defenders in Jordan whose phones were allegedly hacked by parties with links to

⁴ William Christou, “[Jordan Human Rights Defender Hala Al-Ahed Faces Harassment Campaign](#)”, The New Arab, 19 June 2023

⁵ Freedom House, [Jordan](#).

⁶ Human Rights Watch, “[Jordan: Teachers’ Syndicate Closed; Leaders Arrested](#)”, 30 July 2020.

⁷ Mohammad Abu Rumman, “[Protest, Reform and Reaction: Jordan’s Muslim Brotherhood Searches for a ‘Safe Passage’](#)”, 16 September 2020.

⁸ Front Line Defenders, [Jordanian Teachers’ Syndicate \(JTS\)](#)

⁹ Front Line Defenders, [Hala Ahed Deeb](#).

the Jordanian government.¹⁰

Hala, who is not a member of the Brotherhood, holds progressive and different points of view from those that are generally espoused by the Brotherhood and its followers on a number of issues, including women's rights and feminism. But her views also stand at odds with a significant percentage of the Jordanian population, who have more conservative beliefs.

These factors contributed significantly to creating a large constituency of people who were eager to attack Ahed at the first opportunity, which presented itself at the time when Ahel posted the invitation to the event on social media. Ahed's activism and outspoken criticism of government policies had already made her a target of harassment and hate speech from certain segments of society, which can be attributed to deep-rooted political divisions and societal tensions in Jordan.

Methodology

JOSA's monitoring of the hate speech campaign against Ahed ran from the time of the Ahel organisation's post announcing the session on feminism, on 13 June 2023, through 28 August 2023.

Upon identifying the hate speech campaign targeting Ahed, JOSA researchers initiated the manual collection of URLs containing hate speech directed at women in general, and specifically at her. They successfully gathered 79 posts from X (formerly Twitter), although the lack of access to X's application programming interface (API)

impeded the research somewhat.

Following this data-collection phase, the team proceeded to the extraction process, which involved extracting all comments and quote tweets from the posts into a dataset format. To facilitate this extraction process, the researchers utilised the tool exportcomments.com.

The extraction process was not, however, without its challenges. Specifically, the tool encountered difficulties in extracting data from 29 post URLs on X, due to limitations imposed by the X API. Despite these setbacks, the team was able to obtain 850 extracted datasets, encompassing posts, tweets, comments, and quote tweets, all of which were subsequently subjected to annotation by JOSA's annotators.

Prior to beginning the data annotation and classification process, the team cleaned the data, retaining only the necessary information. This contained platform information, post URLs, comment URLs, post IDs, comment IDs, comment content, and the date and time of comments.

JOSA completed the data annotation process in partnership with the TAMAM coalition, and the process included two major steps – annotation and data validation. Nonetheless, before beginning the annotation, the team examined numerous areas, including classification identification, definitions, annotator selection, and annotator training.

JOSA and the TAMAM coalition collaborated to build a clear mind map

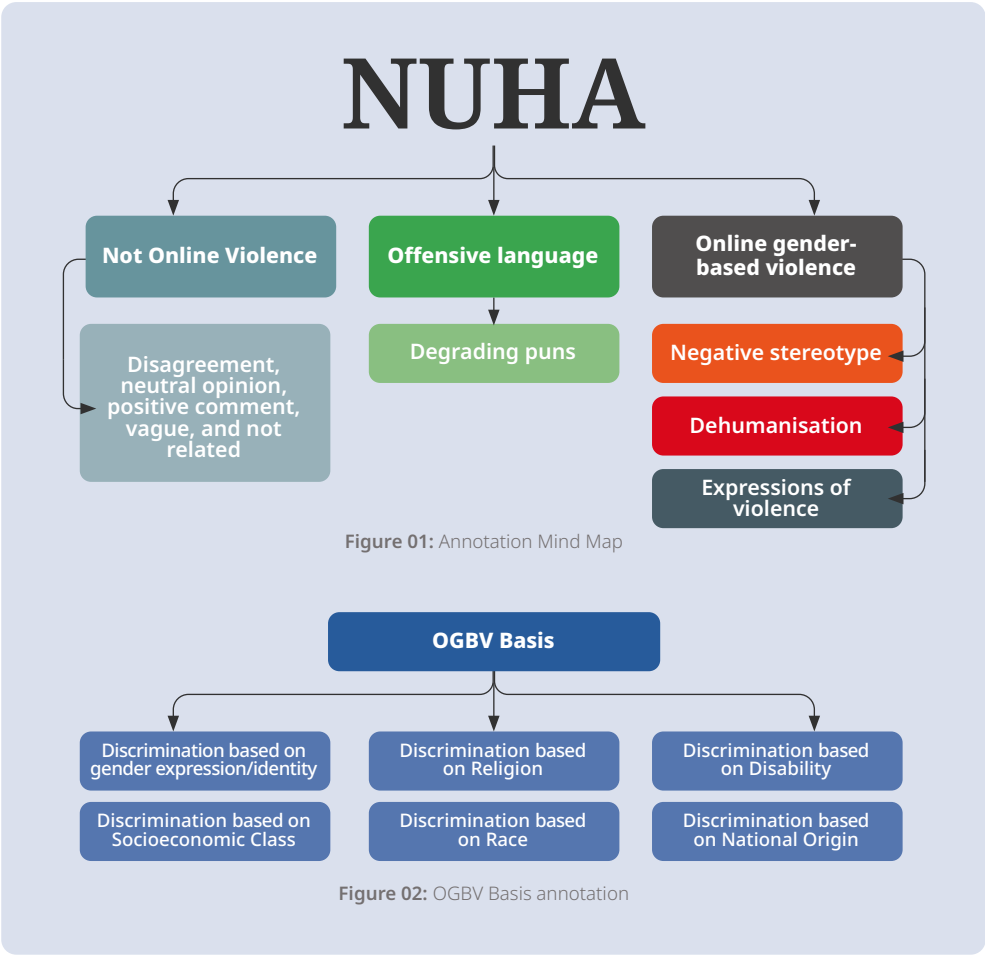
¹⁰ Mohammed Al-Maskati, Bill Marczak, Siena Anstis, and Ron Deibert, "Peace through Pegasus: Jordanian Human Rights Defenders and Journalists Hacked with Pegasus Spyware", The Citizen Lab.

¹¹ Paasch-Colberg, Sünje, et al. "Insults, Criminalisation, and Calls for Violence: Forms of Hate Speech and Offensive Language in German User Comments on Immigration." *Cyberhate in the Context of Migrations*, 2022, pp. 63–137.

based on the work of Monnier et al. in, *“Cyberhate in the Context of Migrations”*.¹¹

The mind map included three classification layers: 1) Not Online Violence, comprising disagreement, positive comments, neutral opinions and not applicable and vague and not related comments; 2) offensive language, containing degrading puns, 3) Online Gender Based Violence (OGBV), which consists of negative stereotypes, Dehumanisation, and expressions of violence. The figures below – Figure 1. Mind map and Figure 2. OGBV Basis – show all three layers that were examined throughout the dataset annotation. In addition, precise definitions for classes and subclassifications were identified, to minimise any confusion during the annotation process (see tables 1 and 2 below for definitions).

After the first part of the annotation (Figure 1), annotators then choose the basis of the hate speech detected, which allowed JOSA researchers to understand the origins and basis of the hate speech (see Figure 2, below).



Hate Speech Terminologies

TERM	DEFINITIONS
Negative stereotyping	The attribution of negatively connoted characteristics, roles, or behaviours to a whole group or to individuals on the basis of their group membership. Verbal attacks (including harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal).
Dehumanisation	Statements that equate or compare humans to inanimate things, animals, or non-human beings, or characterise humans as savage or animalistic.
Expressions of violence	Statements that justify, incite, or threaten physical violence against or the killing of an individual or a group. Calls for exclusion or segregation made against groups can be included.

Table 01: Classification of Hate Speech

Offensive Language Terminologies

CATEGORY	DEFINITIONS
Insults and slurs	A gross indignity – an instance of insolent or contemptuous speech or conduct (Merriam Webster)
Degrading metaphors	Metaphors that contain degrading or humiliating language. The meaning of degrading is causing or associated with a low, destitute, or demoralised state – causing someone to be or feel degraded. (Merriam Webster) For example: Statements that put immigrants in lower positions, and describing them with degrading examples
Degrading puns	Using language of degradation (degrading is causing or associating with a low, destitute, or demoralised state – causing someone to be or feel degraded) in sarcastic speech. For example: The whole text is a pun, but it contains degrading words.

Table 02: Classification of Offensive Language

Furthermore, JOSA adopted the intensity scale shown below, in Figure 03, that was used by other regional partners within the “Words Matter” project, as a reference to analyse the intensity of identified hate speech for analysis, findings, and reporting purposes. The intensity scale was initially measured based on the Social Science Research Council SSRC paper on “Classifying and Identifying the Intensity of Hate Speech”, Nov, 2020.¹²







Color	Title	Description	Examples
	1 Disagreement	Rhetoric includes disagreeing at the idea/belief level. Responses include challenging claims, ideas, beliefs, or trying to change their view.	False, incorrect, wrong, challenge, persuade, change minds
	2 Negative Actions	Rhetoric includes negative nonviolent actions associated with the group. Responses include nonviolent actions including metaphors.	Threatened, stole, outrageous act, poor treatment, alienate
	3 Negative Character	Rhetoric includes nonviolent characterizations and insults. There are no responses for #3.	Stupid, thief, aggressor, fake, crazy
	4 Demonizing and Dehumanizing	Rhetoric includes subhuman and superhuman characteristics. There are no responses for #4.	Rat, monkey, Nazi, demon, cancer, monster
	5 Violence	Rhetoric includes infliction of physical harm or metaphoric/ aspirational physical harm or death. Responses include calls for literal violence or metaphoric/aspirational physical harm or death.	Punched, raped, starved, torturing, mugging
	6 Death	Rhetoric includes literal killing by group. Responses include the literal death/elimination of a group.	Killed, annihilate, destroy

Figure 03: Intensity scale

¹² Bahador, Babak. “Classifying and Identifying the Intensity of Hate Speech.” Social Science Research Council, Social Science Research Council, 17 Nov. 2020, items.ssrc.org/disinformation-democracy-and-conflict-prevention/classifying-and-identifying-the-intensity-of-hate-speech/. Accessed 20 Nov. 2023.

Content Analysis

Statistics and findings:

The gender-based hate speech campaign against Ahed became clear following the Ahel organisation's invitation post on X (Screenshot 2) on 13 June 2023. The post had 1.6 million views, 452 comments on Facebook, and 394 retweets and quote tweets.



Screenshot 2: The Ahel organisation post on X, announcing the event on feminism

Date of post: 13 June 2023

Post text in English: "Residing in Jordan and interested in learning the principles of feminism?"

Now is your opportunity to apply to attend a group of sessions facilitated by Professor Hala Ahed, which carry various titles such as "tyranny", "authoritarianism", "justice", and many others.

Hurry to register through <https://forms.gle/EHci7LGtZzq3SZCG6>

Seats are limited, and priority is given to those who register first!

#Ahel"

Link of post: [Here](#).

The annotation was completed using the Label Studio annotation tool, relying on the mind map. The annotated data sample reflects the findings of online hate speech against Ahed, (See Chart 1, below). Chart 1, reflects the first layer of annotation as, out of the 850 pieces of content extracted, 329 were annotated as hate speech (38.7 per cent), and 521 annotated as non-hate speech (61.3 per cent).

The 61 per cent of the content annotated as non-hate speech was distributed as shown in Chart 2. Thus, 36.5 per cent of this content targeting was annotated as disagreement, which shows that the majority of the non-hate speech content disagree with Ahed's views, without containing any insults or toxic words towards her character or what she represents. The other 24.8 per cent is distributed between vague and not-related content, positive comments, not applicable (emojis or icons), and neutral opinion.

Comments Annotation Result
Hate Speech vs Non-Hate Speech

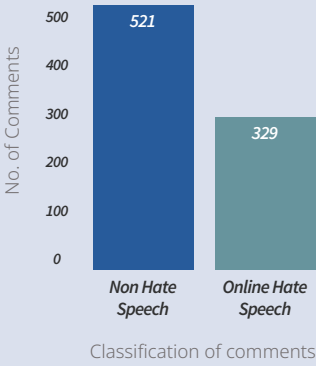


Chart 01: Annotation result hate speech vs. non-hate speech

Sentiment Overall
percentage distribution

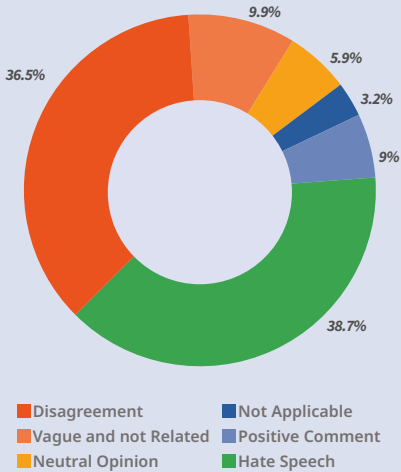


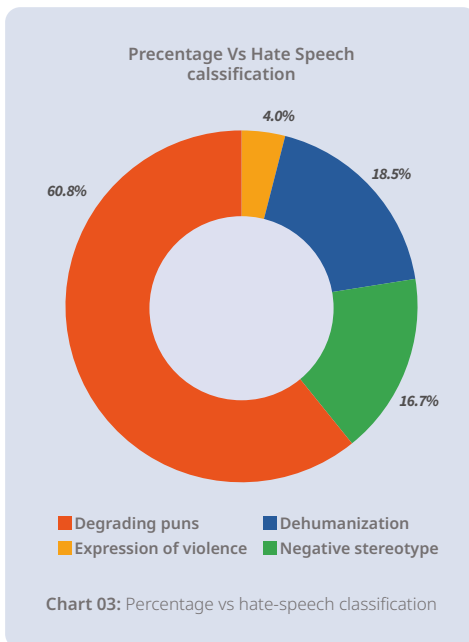
Chart 02: Sentiment Overall Percentage Distribution

A close look at the subclassifications of identified hate speech of the sample in Chart 3, below, reveals that

- 60.8 per cent of the hate speech content against Ahed detected comprised degrading puns, including bullying and insulting content, as well as degrading ironic and sarcastic content;
- 16.7 per cent of the hate speech posts detected comprised negative stereotypes, including hateful misogynist content and sexist stereotypes, as well as negative comments on Ahed's physical appearance;
- 18.5 per cent of hate the speech detected involved the dehumanisation and demonisation of Ahed's character, of feminism, and of women rights organisations, such as the Ahel organisation, that were also targeted in the hate speech campaign. The dehumanisation category also includes accusations of implementing a Western agenda in Jordanian society, due to foreign funding, and making accusations that women rights organisations and women rights activists such Hala Ahed are demons. According to the intensity scale referred to in the methodology (see Figure 3 - intensity scale), 18.5 per cent of the content is classified as of 4th degree intensity, and classified as dangerous.
- 4 per cent of the hate speech content comprised violent expressions, including death threats, threats of other physical harm, and rape threats and sexual harassment.

The significance of this can be seen when looking at a recent JOSA study for the “Words Matter” network’s 3rd Regional Report, where, of a sample of 37,000 comments extracted from 260 posts in an annotation process that was carried out by monitoring the social media accounts of 56 highly influential women in Jordan, 208 of these were annotated as expressions of violence, which is approximately 1.75 per cent, making the 4 per cent figure for the comments aimed at Ahed very serious and a matter for real concern.

- From a comparative point of view, an increase from 1.75 per cent in September 2022 to 4 per cent in May 2023 a very significant, especially when the increase falls under the more violent and extreme categories, as explained above.



Textual Analysis:

Following the data annotation cleaning and analysis, the team put together a lexicon of most-used words in the hate speech campaign against Ahed, and identified them, respectively, as “Overall lexicon”, “Hate speech lexicon”, and “Hate speech unique” (which includes unique words annotated as hate speech).

Analysis revealed that the hate speech lexicon closely resembles the overall lexicon, with a notable emphasis on words like “feminism” (النسوية), “Islam” (الإسلام), “religion” (الدين), and “omen” (النساء). An important discovery by JOSA, however, involved exploring unique terms within the hate speech category. This exploration unveiled the propagation of slurs and insults in the hate speech campaign, connecting conspiracy theories and accusations through terms such as “the devil” (الشيطان) and “enemy” (العدو). Additionally, the motive behind the hate speech was identified, characterised by the intent to protect religion from what is perceived as the “feminist agenda”. This motive is evident in phrases like “O Allah” (يا الله) and “our prophet Mohammad” (نبينا محمد).

Through this analysis, we observe a significant contrast between both hate speech and non-hate speech language by examining their unique lexicons. The non-hate unique lexicon is characterised by more positive and supportive discourse, featuring terms such as “support” (الدعم), “respect” (الاحترام) and “the teacher” (المعلمة) using she/her pronouns, which Jordanians use to show respect to an educated woman). These terms bear no resemblance to the words found in the hate speech-unique lexicon.

Top 10 occurring words

ARABIC WORD	ENGLISH TRANSLATION	OCCURRENCE OVERALL	OCCURRENCE IN HATE SPEECH TWEETS	OCCURRENCE IN NON-HATE SPEECH TWEET
هالة/ هاله	Hala	637	268	369
ديب/ عاهد	Deeb (Hala Ahed family name)	601	247	354
المرأة/ المراه امراة/ النساء/ نساء	Woman	250	43	207
النسوية	Feminism	244	92	152
الله	Allah	195	98	97
الاسلام/ اسلامية الاسلافي... الخ	Islam	150	47	103
الدين	Religion	98	16	66
المجتمع/مجتمع مجتمعنا .. الخ	Society/ Our society/ Your Society/ etc.	80	15	65
😊	😊	66	32	34
النسويات/ نسويات	Feminists	29	18	11
الغرب/ مستغرب	West/ Westernised			

Table 03: Top 10 Occurring words (Overall, hate speech content, non-hate speech content).

Table 4, below – Top 10 words in the hate speech-unique lexicon – presents the hate speech-unique lexicon and top words that were used in a hate speech-annotated content. It is worth highlighting that the most often occurring words mentioned in table 1 are different from those identified in table 2.

Top 10 words in the hate speech-unique lexicon

ARABIC WORD	ENGLISH TRANSLATION	OCCURRENCE
مطالب	Demands	11
كافر / كافرة / كافره / كفار / الخ	Faithless/Kafir	10
جهل	Ignorance	8
لواط / شنود / شانة / الخ	Gay/faggot/homosexual	8
نبينا محمد	Our prophet Mohammad	7
🚫	🚫	6
عدو	Enemy	6
الشيطان	The devil	6
يا الله	O Allah	5
عاهرة	Whore	4
تفسير	Explanation	4

Table 04 clearly shows how slurs and insults were perpetrated in the hate speech campaign and linked with conspiracy theories and accusations through words such as: “the devil” and “enemy”. It also reveals the motive, which is protecting religion from the feminist agenda, through terms such as “O Allah” and “our prophet Mohammad”.

In table 5, above, we can see the significant shift between both hate speech and non-hate speech language by looking at the unique lexicons; the non-hate speech-unique lexicon represents a more positive or supportive discourse, using terms such as “support”, “respect”, and “the teacher” (using she/her pronouns Jordanians use to show respect to an educated woman). These words have no resemblance to the words in the hate-speech-unique lexicon.



Screenshot 5: Comment on the Ahel organisation post on X

Date of post: 14 June 2023

Post text in English: *"Feminism is not against the Islamic religion only, it is also against the teachings of the Lord of Islam, the Qur'an, and the Sunnah.*

We need to explain so you can understand"

Link of post: [Here](#).

As shown in the screenshots, there were three main trends in the attack on Ahed:

1. Sexist and homophobic speech;
2. Demonisation of feminists and feminism in general; and
3. Hate speech content with religious basis.

Conclusion and recommendations

The case study of the hate speech campaign against Hala Ahed sheds light on the complex challenges faced by individuals advocating for human rights, gender equality, and freedom of expression in Jordan. Despite her acknowledged work and international recognition, Ahed has encountered various forms of violence and threats, including online hate speech campaigns and the infection of her phone with spyware.

The case study of Hala Ahed calls for increased awareness, support, and action to combat online hate speech and to promote the protection of fundamental human rights, gender equality, and freedom of expression. Ahed's resilience in the face of adversity serves as a testament to the importance of advocating for justice and equality in challenging environments.

Additionally, this case study offers valuable insights into the dynamics of public engagement and online campaigns against individuals like Ahed. It is crucial to recognise that Ahed is neither the first nor the only woman who has been targeted in such a manner. Similar cases can be found both within the region and worldwide, where women human rights defenders and activists face intimidation, harassment, and hate speech due to their advocacy work.

The study underscores the importance of addressing these issues comprehensively, not only to support and protect individuals like Ahed, but also to contribute to a broader understanding of the challenges faced by women and human rights defenders globally.

To prevent targeted internet campaigns against women in public spaces, a multidimensional solution, combining efforts by tech corporations, government agencies, civil society, and individual users is required. Here are some suggestions about how to deal with this problem:

1. Stricter Regulation:

- Governments must enact and rigorously enforce anti-online harassment and hate speech laws.
- Governments should ensure that existing legislation comprehensively addresses internet harassment, with severe penalties to deter offenders, recognising the challenges faced by activists like Hala Ahed.
- Governments should strengthen legal protections for organisations advocating for human rights and liberties, to prevent unwarranted interference.

2. Platform Responsibility:

- Platforms should streamline and expedite their hate speech reporting processes, treating reports with urgency.
- Platforms should ensure prompt and effective actions in response to reported instances of online harassment.

3. User Education:

- The authorities should organise the conduct of comprehensive user education programmes on responsible online behavior and the repercussions of harassment.
- The relevant bodies should promote digital literacy and awareness initiatives, particularly targeting young users to identify and combat online abuse.

- The relevant bodies should develop targeted programmes that address the specific challenges faced by women activists, countering societal prejudices and fostering empathy.

4. Victim Assistance:

- The authorities should create helplines and support services for victims of online harassment.
- Tech businesses could provide tools for mental health support to those who have been harassed online.
- The authorities should collaborate with tech experts to establish protocols that safeguard individuals' digital privacy, particularly those targeted for their advocacy work.
- Legal assistance and representation should be provided to activists facing unwarranted persecution, reinforcing the importance of upholding freedom of expression.

5. Media Literacy:

- Governments should introduce media literacy programmes in schools, to assist students in critically analysing and evaluating online content.
- Media organisations should be encouraged to portray women in diverse, empowering roles in their content.

6. Accountability and Transparency:

- Regular updates on tech companies' efforts to fight online harassment, including the number of reported occurrences and actions taken, should be published.
- Third-party audits should be established to ensure that content-moderation policies and methods are transparent.
- Media syndicates, CSOs, and independent regulatory state institutions should promote media integrity and unbiased coverage to counteract the divisive narratives that escalate tensions between government critics and supporters.
- Multi-disciplinary national and regional stakeholders should engage with international organisations to ensure that human rights violations, including online harassment, are thoroughly investigated and addressed.
- Governments should hold platforms accountable to swiftly respond to and mitigate instances of online gender-based hate speech, ensuring a safer digital environment for activists.

7. Research and Collaboration:

- Academic institutions and public higher education regulatory bodies should plan for and facilitate collaboration among

tech companies, academics, and non-governmental organisations to better understand the dynamics of online abuse and to create effective remedies.

- Governments should allocate funding for the conduct of research on the psychological and social aspects of online harassment, so as to inform interventions.

References

1. **"Hala Ahed Profile"**, Front line Defenders, at: <https://www.frontlinedefenders.org/en/profile/hala-ahed-deeb>
2. **"Jordanian lawyer Hala Al-Ahed wins Front Line Defenders award"**, News, Jordan News, May 27 2023, at: <https://www.jordannews.jo/Section-109/News/Jordanian-lawyer-Hala-Al-Ahed-wins-Front-Line-Defenders-award-28873>
3. **Marwa Fatafta, "Women human rights defenders speak out about Pegasus Attacks"**, Post, Access Now, Jan 17 2022, at: <https://www.accessnow.org/women-human-rights-defenders-pegasus-attacks-bahrain-jordan/>
4. **Willam Christou, "Jordan human rights defender Hala Al-Ahed faces harassment campaign"**, News, The New Arab, June 19 2023, at: <https://www.newarab.com/news/jordan-human-rights-defender-faces-harassment-campaign>
5. **"Jordan" Freedom In The World 2023**, Freedom House, at: <https://freedomhouse.org/country/jordan/freedom-world/2023>
6. **"Jordan: Teachers' Syndicate Closed; Leaders Arrested"**, News, Human Rights Watch, July 30 2020, at: <https://www.hrw.org/news/2020/07/30/jordan-teachers-syndicate-closed-leaders-arrested>
7. **Mohammad Abu Rumman, "Protest, reform and reaction: Jordan's Muslim Brotherhood searches for a 'safe passage.'"**, Opinion, Middle East Eye, September 16 2020, at: <https://www.middleeasteye.net/opinion/what-does-future-hold-jordans-muslim-brotherhood>

8. **"Jordanian Teachers' Syndicate (JTS) Profile"**, Front Line Defenders, at: <https://www.frontlinedefenders.org/en/organization/jordanian-teachers%E2%80%99syndicate-jts>
9. **Hala Ahed Profile**, Front Line Defenders, at: <https://www.frontlinedefenders.org/en/profile/hala-ahed-deeb>
10. **Mohammad Al-Maskati, Bill Marczak, Siena Ansits, and Ron Deibert, "Peace through Pegasus: Jordanian Human Rights Defenders and Journalists Hacked with Pegasus Spyware"**, Research, The Citizen Lab, April 5 2022, at: <https://citizenlab.ca/2022/04/peace-through-pegasus-jordanian-human-rights-defenders-and-journalists-hacked-with-pegasus-spyware/>
11. **Paasch-Colberg, Sünje, et al. "Insults, Criminalisation, and Calls for Violence: Forms of Hate Speech and Offensive Language in German User Comments on Immigration."** Cyberhate in the Context of Migrations, 2022, at: https://link.springer.com/chapter/10.1007/978-3-030-92103-3_6.
12. **Babak Bahador, "Classifying and Identifying the Intensity of Hate Speech"**, Research, Items, SSRN, Nov 17 2020, at: <https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/classifying-and-identifying-the-intensity-of-hate-speech/>
13. **Rebecca Godard, Susan Holtzman, "The Multidimensional Lexicon of Emojis: A New Tool to Assess the Emotional Content of Emojis"**, Frontiers, 10 June 2022, at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2022.921388/full>

Third Report

Detecting Online Gender-Based Violence: The Nuha AI Initiative

Introduction

The Jordan Open Source Association (JOSA) is currently developing an artificial intelligence (AI) model called "Nuha", the Arabic word for "mind" or "brain." The model's main purpose is to assist researchers in identifying online gender-based violence against women in Jordan, particularly on social media platforms.

In this document, JOSA provides an overview of the technical specifications of the AI-powered Nuha tool and its development, covering the model's architecture, the training process, and the model's evaluation. Additionally, the document addresses some of the related ethical considerations, as well as the model's usability, sustainability and limitations, the lessons learned in its development so far, and potential improvements.

The report is split into two sections: one aimed at readers with limited technical understanding, providing

an overview of Nuha and its basic functionalities, and another section tailored for researchers and technical experts seeking in-depth insights into Nuha's technical workings. Both types of readers can find the information they need based on their expertise. Scientifically precise technical terms are used intentionally for accuracy, and their meanings within this report are detailed in the table below for clarity.

Problem Statement and Objectives

As hate speech continues to be a significant concern for human rights in Jordan and the MENA region, with a notable increase in the online targeting of women, research conducted by JOSA indicates that hate speech has been increasingly used as a tool, with attempts to silence and intimidate political activists and human rights defenders accounting for 32.5 per cent of all forms of gender-based hate speech, including 6.3 per cent

using offensive language.

Women in Jordan face various issues in the digital sphere, ranging from cyberbullying to non-consensual sharing of intimate images, leading to actual physical harm and emotional or psychological suffering.¹ The rapid spread of internet access and social media platforms has opened up new channels for communication, but it has also increased online violence and, more specifically, online gender-based violence directed at women.

As addressing online gender-based violence (OGBV) and hate speech is crucial for creating a safer, more inclusive digital environment for women in Jordan and globally, Nuha aims to enhance the protection of women in Jordan, through:

1. Improving the monitoring and reporting of online hate speech, through assisting researchers in identifying OGBV on social media platforms; and
2. Enhancing the knowledge and capacities of at-risk individuals to mitigate and combat online gender-based hate speech, through drawing insights from Nuha's analysis, understanding patterns and the basis of online gender-based hate speech.

Data Collection and Preprocessing

Data collection and processing is a crucial aspect of the machine learning

lifecycle. Consequently, JOSA, with the support of the TAMAM Coalition (the Women's Digital Safety Alliance), in Jordan, has implemented a well-defined process that involves several sub-steps, and technical tools. The process was defined based on in-house technical experiences, desk-research, and a number of meetings and round tables to discuss, form, and finalise the process, as detailed below.

1. Accounts of women in the Jordanian public space:

In collaboration with the TAMAM Coalition, JOSA identified 60 accounts of women in Jordanian public spaces and 20 online campaigns for women's rights, including civil and political rights. In order to generate a representative sample, the identified accounts were selected based on specific selection criteria, including women from diverse backgrounds, ranging from public figures to politicians and human rights defenders, who have an online presence, and are subject — or are potentially subject — to online gender-based hate speech.

2. Account monitoring:

The monitoring of the 60 identified accounts was divided into five rounds, primarily linked to women's online activity from the previous two years and any recent online presence. Some of those accounts were included in multiple monitoring rounds due to new instances of online violence that occurred at a later stage. JOSA solely

¹ The Jordanian National Commission for Women, "Violence against Women in the Public and Political Spheres in Jordan 2022" (in Arabic), 2022.

used Crowdtangle ² for content searches, and systematically recorded relevant post URLs in a spreadsheet for analysis and processing.

3. Data collection:

The overall dataset, comprising more than 85,000 comments from 419 posts that included instances of online violence, covered the 60 identified accounts, as detailed in table 1.

JOSA used the Export Comment ³ tool for the data collection, and securely stored the data internally, with restricted access. This precaution was taken due to the potential presence of sensitive information (see the Ethical Considerations section).

DATA COLLECTION ROUND	DURATION	PLATFORM	NO. OF ACCOUNTS	NO. OF POSTS	NO. OF COMMENTS
1	Sep - Dec 2022	Facebook	10	116	25,215
2	Dec 2022 - Feb 2023	Facebook	23	149	15,209
3	Apr - May 2023	Facebook	22	89	28,346
4	Jun - Aug 2023	Facebook & Twitter	1	49	852
5	Aug - Oct 2023	Facebook	6	16	16,967

Table 01: Data Collection Rounds

4. Data cleaning;

The team conducted a data cleaning process before commencing data annotation, retaining only the essential information, including the platform, post URL, comment URL, post ID, and comment ID, as well as the text of the comments/tweets/posts, and the date and time of the comments.

JOSA utilised three main online tools in the data collection and cleaning phases, in order to ease and speed the process:

1. The Crowdtangle tool was used to search for online content on Meta social media platforms within the past two years;
2. The Export Comment tool was used to manually extract Facebook comments and tweets, post-by-post; and

² Crowdtangle, "A Tool from Meta to Help Follow, Analyze, and Report on What's Happening Across Social Media".

³ www.exportcomments.com

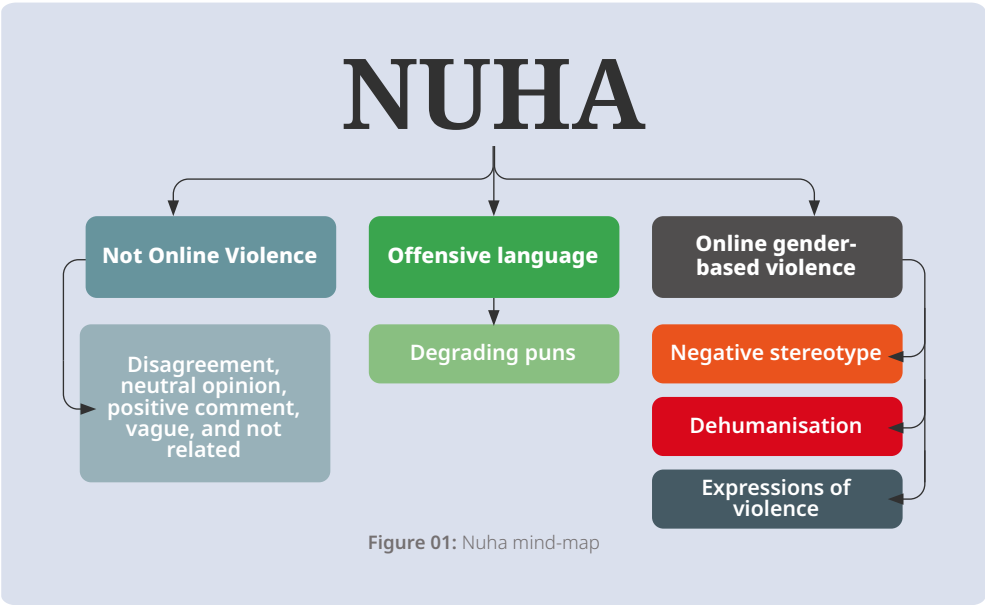
3. Label Studio ⁴ was used to store the data collected for the dataset annotation process (see the Dataset Annotation section for further information).

Dataset Annotation

The dataset annotation mainly focuses on three sub-steps, including the preparation, annotation, and data validation, as clarified below.

Preparation and Dataset Annotation Tools

Working alongside DRI and the TAMAM coalition, JOSA developed a mind-map detailing the annotation classifications. These are presented in three tiers: 1) determining whether it is online gender-based hate speech; 2) classifying the type of gender-based hate speech; and 3) identifying the basis of the hate speech, based on the categories of gender expression/identity, religion, disability, socioeconomic class, race, or national origin, as illustrated in Figure 1.



To differentiate hate speech from other expression, JOSA created a second layer of the mind-map, based on the work of Monnier et al. in *"Cyberhate in the Context of Migrations"*. ⁵ Ambiguous content, like irony and sarcasm, was categorised as "degrading puns", under offensive language. (For detailed definitions, see the Nuha Glossary, in the appendix).

⁴ Label Studio, "Open Source Data Labeling Platform".

⁵ Paasch-Colberg, Sünje, et al. "Insults, Criminalisation, and Calls for Violence: Forms of Hate Speech and Offensive Language in German User Comments on Immigration.", op. cit., note 6.

At the same time, OGBV is categorised into three levels of color-coded intensity (orange, red, black). The "Negative Stereotype" category is the mildest in orange as seen in (figure 1), and includes misogyny, sexism, insults, bullying, and remarks about physical appearance. The "Dehumanisation" category is of a medium intensity (in red), comprising accusations and demonisation. Lastly, the "Expression of Violence" category is the most severe (appears in black), including sexual harassment and threats of rape, physical harm or death, and information security threats. The intensity scale was inspired by the study of Classifying and Identifying the Intensity of Hate Speech (Babak Bahador –2020), as shown following the same color-coded intensity scale introduced) in Figure 2.







Color	Title	Description	Examples
	1 Disagreement	Rhetoric includes disagreeing at the idea/belief level. Responses include challenging claims, ideas, beliefs, or trying to change their view.	False, incorrect, wrong, challenge, persuade, change minds
	2 Negative Actions	Rhetoric includes negative nonviolent actions associated with the group. Responses include nonviolent actions including metaphors.	Threatened, stole, outrageous act, poor treatment, alienate
	3 Negative Character	Rhetoric includes nonviolent characterizations and insults. There are no responses for #3.	Stupid, thief, aggressor, fake, crazy
	4 Demonizing and Dehumanizing	Rhetoric includes subhuman and superhuman characteristics. There are no responses for #4.	Rat, monkey, Nazi, demon, cancer, monster
	5 Violence	Rhetoric includes infliction of physical harm or metaphoric/ aspirational physical harm or death. Responses include calls for literal violence or metaphoric/aspirational physical harm or death.	Punched, raped, starved, torturing, mugging
	6 Death	Rhetoric includes literal killing by group. Responses include the literal death/elimination of a group.	Killed, annihilate, destroy

Figure 02: Classifying and Identifying the Intensity of Hate Speech (Babak Bahador –2020)

During our workshop, five TAMAM coalition experts discussed key OGBV concepts and the mind-map, and agreed on annotation guidelines.⁶ As part of the preparations, the open source data labelling tool Label Studio was used to reflect only the main information, i.e., comment link, comment text, platform, and the three classification layers.

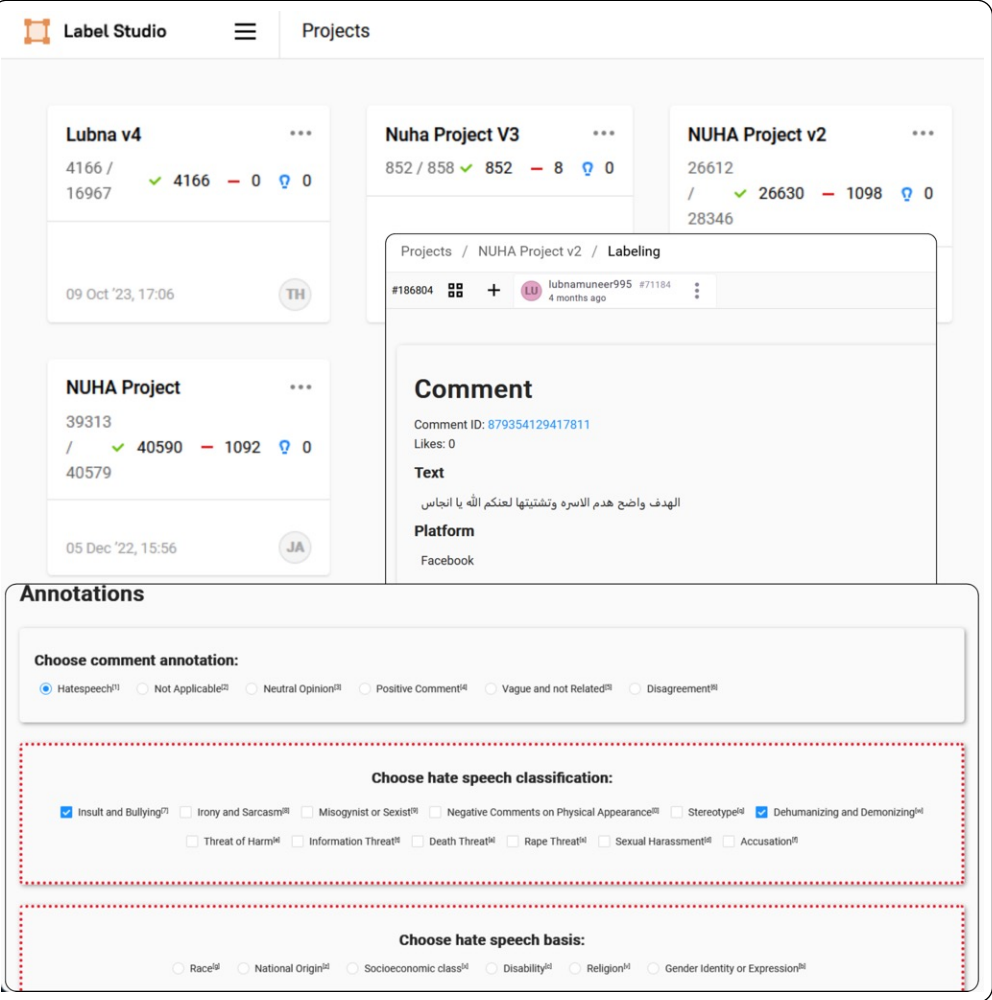


Figure 03: Screenshot of an Example of Label Studio

⁶ Guidelines are accessed by JOSA's researchers and five pre-selected annotators. Annotation should follow the provided mind-map, and each piece of content can have, at most, three sub-classifications. Only Arabic content is annotated.

Dataset Annotation and validation

JOSA annotated a vast dataset of over 70,000 comments from 419 posts, monitored across 60 women's accounts. Additionally, the team completed human and automated data validation to lower the human-error percentage, including in bias annotation. The data validation took two different shapes:

- **Random Check:** JOSA completed a random annotation check for approximately 10 per cent of the "Round 1 and Round 2" dataset sample, and made direct amendments (corrections or adjustments to the data annotation), as needed.
- **Nuha – Human Check:** The human check phase commences upon achieving a confidence level ⁷ with the human annotation. In cases of opposing annotations with high confidence, i.e., when the model's annotations disagreed with the human annotations and the model was confident in its answers, JOSA made direct corrections to the data to ensure accuracy and quality. Approximately 10 per cent of the overall sample was re-annotated based on the data validation process.

Dataset Analysis

Upon the completion of the annotation and validation, JOSA analysed only the annotation for the first four rounds for research purposes. It is important to emphasise that this dataset represents preliminary findings regarding online gender-based hate speech in Jordan, as depicted in the Figures below.

⁷ Confidence Level: A measure that reveals how sure we can be about the accuracy of the annotations made by the computer model. It is essentially a measure of how much trust we can have in the model's decisions. The higher the confidence level, the more certain we can be that the model's annotations are correct.

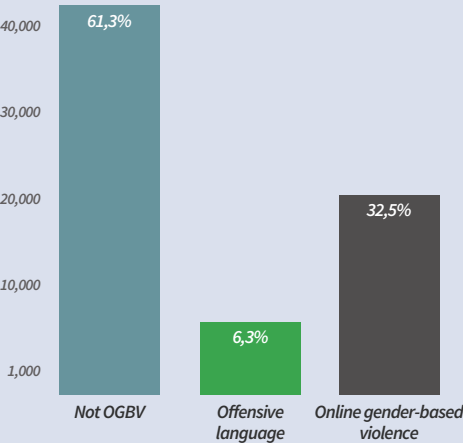


Figure 04: Annotation classification

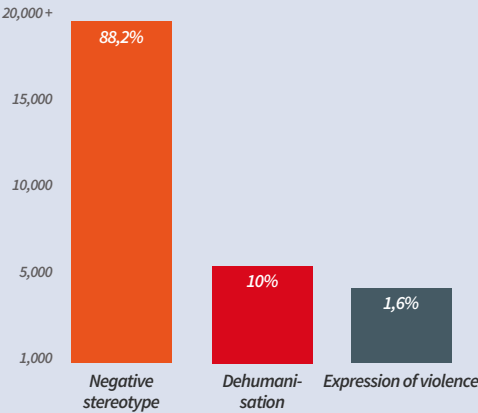


Figure 05: Annotation Hate Speech Subclassification

Model Architecture

Text data is inherently “high-dimensional”, indicating a wealth of diverse words with various stems, resulting in a complex network of interdependencies among words within sentences. This means that words in a single sentence are very dependent on each other. Take, for example: “you are nice”, “you are not nice”, “you might think you are nice” – in Arabic, “أنت لطيف” and “قد تعتقد أنك لطيف”, “أنت لست لطيفا”, respectively. All of these sentences have the word “nice”, or “لطيف”, in them, but within different contexts and, therefore, with different meanings.

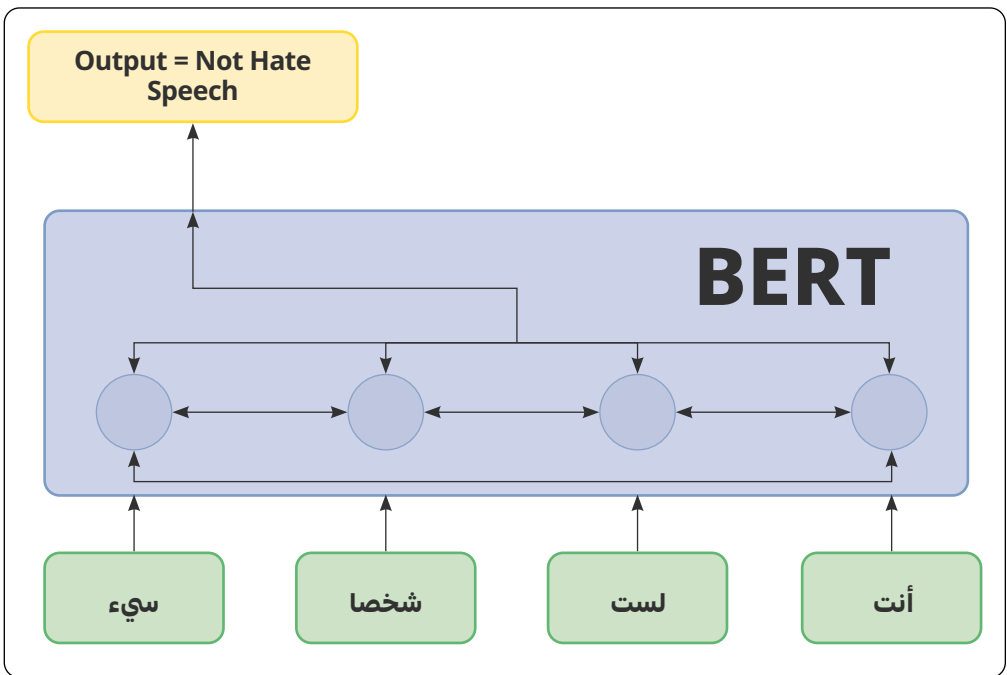


Figure 06: Illustration of BERT

JOSA adopted the **BERT**⁸ transformer model for Nuha’s development. BERT is a deep learning architecture considered revolutionary in its ability to understand intricate text patterns.

BERT undergoes a pre-training phase, where it learns from a vast amount of text data, understanding contextual word relationships by predicting missing words from sentences. After

⁸ Arxiv.org, “BERT: [1810.04805] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Cornell University.

pre-training, BERT can be tailored for tasks like hate-speech detection. It uses its transformer architecture to produce embeddings, which are vector representations of words that capture their meanings and context. By using BERT, JOSA can leverage the model's advanced capabilities to enhance the quality and accuracy of Nuha's text-related development, leading to more effective results in text classification.

In conjunction with BERT, JOSA used AraBERT and other Arabic pre-trained models. These pre-trained models provide the basis for Nuha's Arabic language processing capabilities, contributing significantly to its success.

Central to Nuha is the Masked Language Model (MLM), which uses techniques to improve language comprehension during training. MLM training involves predicting words and understanding context, much like human language comprehension.

The Training Process

During its training phase, JOSA divided the dataset into an 80-10-10 ratio for training, validation, and testing, respectively. The validation set was pivotal in gauging the model's efficacy.

Initially, a baseline model "starting point" was established, and then updated using the training data. Subsequently, a Transformer-based model⁹ was fine-tuned. This model served as the foundation for constructing a classifier,

forming the basis for building Nuha.

Handling imbalanced datasets, where one class or category of sentences prevails over another, was a challenge during model training, making the model less adept at identifying less-common sentences. To address this, JOSA used under-sampling to balance the representation, by reducing the amount of the overrepresented data (content that was not hate speech) to make it more equal with the underrepresented data (hate speech content). This is like giving more attention to the less common group to create a balanced view. In addition, JOSA used data augmentation¹⁰ to boost the less represented class, by slightly modifying existing samples and adding them to the dataset.

Considering the limited volume of data, measures were taken to prevent overfitting (learning too much), where the model memorises examples, rather than grasping underlying patterns. In addition to data augmentation, model size reduction (reducing the capacity of the model, so it is not able to overfit) was deemed essential to provide better results. Multiple experiments were conducted to determine optimal hyperparameters (settings that the individual can adjust to fine-tune how the model learns), including the learning rate, batch size, and weight decay, to optimise the learning process and adapt to Nuha's dataset nature.

⁹ A transformer-based model is a deep learning architecture, initially proposed in 2017, that relies on the parallel multi-head attention mechanism. See: A. Aswani et al. "[Attention Is All You Need](#)", Cornell University, 12 June 2017.

¹⁰ Datacamp, "[A Complete Guide to Data Augmentation](#)", November 2022.

Model Evaluation

JOSA completed regular monitoring of the model's performance, through running a comparison between recently trained data and a newly annotated dataset that was not exposed to the model. This technique allowed JOSA to identify potential shifts in data distribution, by identifying significant changes or differences in the patterns or characteristics of the data. This means checking whether the data looks different or behaves differently from previously monitored trained data. This is important, because such shifts can affect the model's performance; by monitoring for them, JOSA can adapt the model to handle new or changing data effectively.

The model's performance was evaluated by utilising essential metrics, including the F1 score, precision, and recall. The primary metric for evaluating Nuha's sentence recognition and classification was the F1 score, which combines two key elements: precision (the model's accuracy) and recall (model's ability to detect all relevant instances). The score ranges from 0 to 1, with a higher value indicating better model performance.

Limitations

During the training and evaluation processes, JOSA faced challenges including:

- Twitter application programming interface (API): As a result of a recent policy change by X (formerly Twitter), JOSA was not able to gain access to the platform's API. This was an impediment to account monitoring and research components.
- Data imbalance: It is important to ensure a balanced representation of hate speech and non-hate speech, in order for the tool to effectively learn to identify hate speech. Ideally, the tool should be exposed to both hate speech and non-hate speech equally. Based on the annotated dataset, the tool was exposed to a higher percentage of data labelled as non-hate speech than the percentage of data labelled as hate speech. This data imbalance was countered with data resampling, data augmentation, and weighted class loss functions.
- Nuha is best suited for complex, non-linear deep learning models, especially considering the intricate nature of text. These models thrive on larger datasets, but there is a risk of overfitting when working with smaller datasets. To mitigate this risk, we opted for smaller models with reduced learning rates. Learning rates determine the size of the steps a model takes in finding the optimal parameter values. If the rate is too high, the model may overshoot optimal values and fail to

converge properly. On the other hand, if it is too low, the model might take an extended time to converge, or become stuck in suboptimal solutions.

- Noise in the labels (errors, inconsistencies, or inaccuracies in the labelled data used for training and evaluation) was later detected, and was filtered and re-annotated to point out labels that contradict predictions from the model with high confidence.

Deployment

The deployment had three parts: a web client,¹¹ APIs,¹² and the model dataset. Cloud scalability was achieved through automated scaling on cloud computing, ensuring no downtime to users when undergoing deployment. This means that Nuha will be available to users through Nuha API by hosting it on JOSA's servers and systems, making it independent from external API providers and less susceptible to being controlled or restricted by third parties, which ensures that it remains available and accessible to users (0 downtime). Both the web client and the model's target platform are hosted on JOSA's cloud servers.

The web client is a user interface for the Nuha APIs,¹³ while the API client runs the NUHA model hosted on Hugging Face. The API processes user requests to detect hate speech, runs the model, and returns results, which may result in some latency, based on internet speed and data sample size.

¹¹ Web client (browser): A web browser is an application for accessing websites and the Internet. The New Dictionary Of Cultural Literacy: What Every American Needs to Know (3rd ed.) Hirsch, Eric Donald (2002).

¹² API: A way for two or more computer programs to communicate with each other. It is a type of software interface, offering a service to other pieces of software. API Design for C++ Reddy, Martin (2011)

¹³ Open API documentation is available at: Nuha API - Swagger UI.

Some examples of the API's request and responses are shown in the following screenshot:

Nuha API

0.1.0

QAS 3.1

/openapi.json

API to serve ML model for hate-speech classification

default

GET /healthcheck Healthcheck

POST /predict Predict

Classify comments into hatespeech or not.

Parameters

Cancel

Reset

No parameters

Request body required

application/json

```
[
  {
    "comment": "تقول بنت نيل مش شايج مرسه",
    "post": "إحنا متحمسين لشوف عمل آرني جديد على تيلتكس تجربة انشاء الله نمشي اللي صار زمان مع مفرجة أرنية المحتمل انها تيهو المشافد... حكون واقعين بالشفد... بالانتظار"
  },
  {
    "comment": "ارج بكوكي محترمت بالتمثيل بستيوهات تيلتكس عكن حقيقتهم عاتوانك يستيوهات",
    "post": "إحنا متحمسين لشوف عمل آرني جديد على تيلتكس تجربة انشاء الله نمشي اللي صار زمان مع مفرجة أرنية المحتمل انها تيهو المشافد... حكون واقعين بالشفد... بالانتظار"
  },
  {
    "comment": "بعدا عن الموضوع الشفه صي تشكن ان شاء الله مش انشاء الله",
    "post": "إحنا متحمسين لشوف عمل آرني جديد على تيلتكس تجربة انشاء الله نمشي اللي صار زمان مع مفرجة أرنية المحتمل انها تيهو المشافد... حكون واقعين بالشفد... بالانتظار"
  },
  {
    "comment": "بطوان مش صبح لكلي هاه",
    "post": "إحنا متحمسين لشوف عمل آرني جديد على تيلتكس تجربة انشاء الله نمشي اللي صار زمان مع مفرجة أرنية المحتمل انها تيهو المشافد... حكون واقعين بالشفد... بالانتظار"
  }
]
```

Execute

Clear

Figure 07: A screenshot of Nuha's API request Body

Response:

Response body

```
[
  {
    "label": "hate-speech",
    "score": 0.9527454376220703,
    "model_version": "v1.0",
    "comment": "نقول بنت ايل من طلاب مدرسه",
    "post": "إخا متحمسين نشوف عمل أردني جديد على نيتفكس تجربة انشاء الله نمحي اللي صار زمان مع مخرجة اردنية المفضل انها تبهر المشاهد.. حكون واقعين بالتقد... بالان"
  },
  {
    "label": "hate-speech",
    "score": 0.7891517281532288,
    "model_version": "v1.0",
    "comment": "روح يكون مخزومات بقتنيل بسكيوهات نتفكس عكس حقيقتهم علانوع بسكيوهات",
    "post": "إخا متحمسين نشوف عمل أردني جديد على نيتفكس تجربة انشاء الله نمحي اللي صار زمان مع مخرجة اردنية المفضل انها تبهر المشاهد.. حكون واقعين بالتقد... بالان"
  },
  {
    "label": "non-hate-speech",
    "score": 0.873494029045105,
    "model_version": "v1.0",
    "comment": "بعيدا عن الموضوع التلقه من تنكتب ان شاء الله من انشاء الله",
    "post": "إخا متحمسين نشوف عمل أردني جديد على نيتفكس تجربة انشاء الله نمحي اللي صار زمان مع مخرجة اردنية المفضل انها تبهر المشاهد.. حكون واقعين بالتقد... بالان"
  },
  {
    "label": "non-hate-speech",
    "score": 0.9784269332885742,
    "model_version": "v1.0",
    "comment": "يخوان مثل صبح النكي هاد",
    "post": "إخا متحمسين نشوف عمل أردني جديد على نيتفكس تجربة انشاء الله نمحي اللي صار زمان مع مخرجة اردنية المفضل انها تبهر المشاهد.. حكون واقعين بالتقد... بالان"
  }
]
```

Download

Figure 08: A screenshot of Nuha's API Response

The main approach JOSA is adopting for maintaining the deployment environment is open sourcing the deployment environment, as developers and organisations who are interested in the project can contribute to the source code. In addition, contributors can run these projects, using a provided [Docker](#) image on their machines or servers.¹⁴

¹⁴ Detailed instructions for running the deployment environment are uploaded to JOSA's GitHub repos, as referenced below:

- [NUHA's web client source code](#).
- [NUHA's API client source code](#).
- [NUHA's Huggingface Dataset model](#).

The figure below represents a graphical illustration of the system's components.

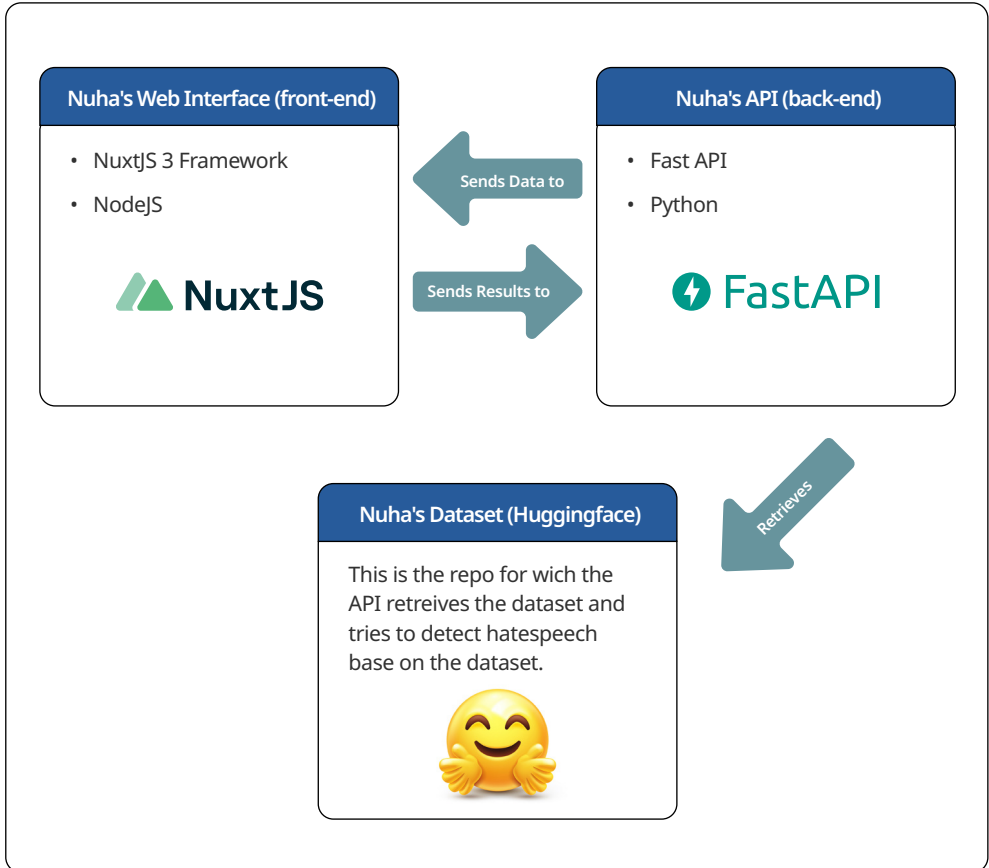


Figure 09: A screenshot of Nuha's project architecture

Ethical Considerations

JOSA's commitment to ethical AI principles is maintained in every phase of the NUHA project, and especially its deployment. Key ethical considerations include:

- 1. User Privacy and Data Security:** Nuha prioritises user privacy and data security. To ensure this, data is anonymised and encrypted, maintaining the confidentiality of user information. JOSA's policies strictly forbid the storing of any personal data without the user's consent. Even post-launch, as users

upload samples to the tool, they must follow a format ¹⁵ that allows data processing without processing personal data.

- Bias Recognition and Mitigation:** JOSA acknowledges the potential for bias at various project stages, from data annotation to model predictions. To counteract – or, at least, reduce – this bias, several strategies were employed. These include careful selection of annotators with solid experience and knowledge of OGBV, rigorous quality checks, and under-sampling to ensure a balanced representation of sentences. Regular audits were also conducted to detect and address any emerging biases.

Usability and sustainability

The AI model will be publicly available to researchers, civil society organisations, and human rights organisations, at <http://nuha.josa.ngo>. The Nuha landing page is designed for simplicity and user-friendliness.

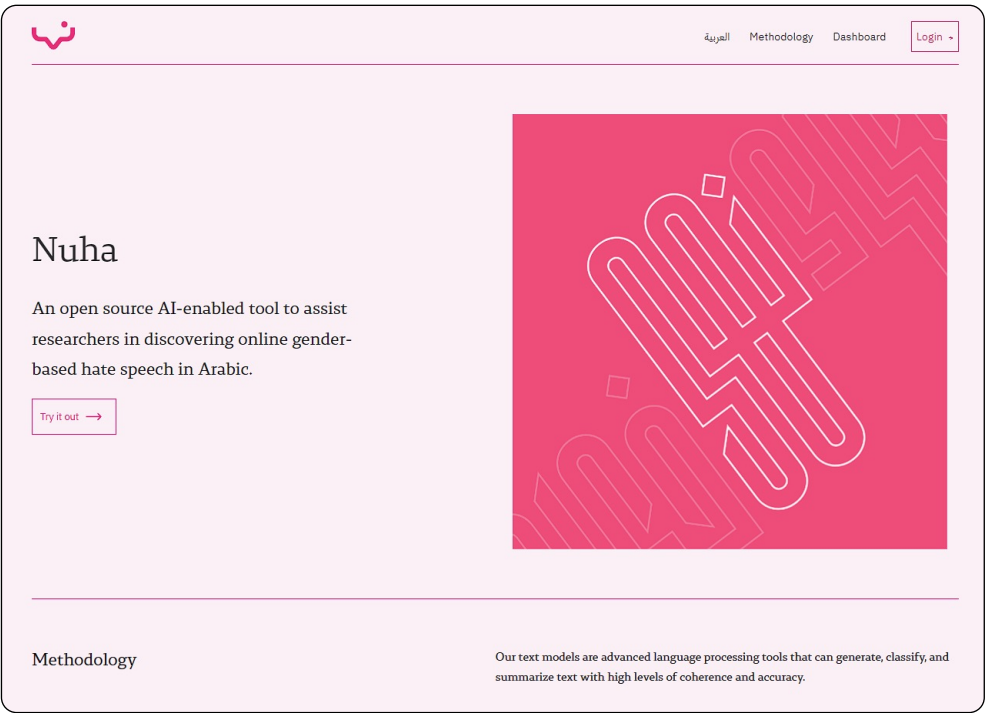


Figure 10: A screenshot of Nuha's homepage

¹⁵ Format includes minimal data without any personal data, including the platform, post link, comment link, timestamp, text comment/ tweet, and number of likes. Format available at <https://nuha.josa.ngo/>

Upon visiting the page, users will find a straightforward interface that encourages interaction with the AI model. Interaction occurs after the user logs into NUHA, where they can either upload their sample data (following a specific format) or enter a comment suspected of containing hate speech.

Currently, users can directly upload sample data using a form. After uploading the file, the system will alert the user about any errors. Processed results are displayed on the same page. Alternatively, users can input a text directly, and NUHA will evaluate the text for potential hate speech.

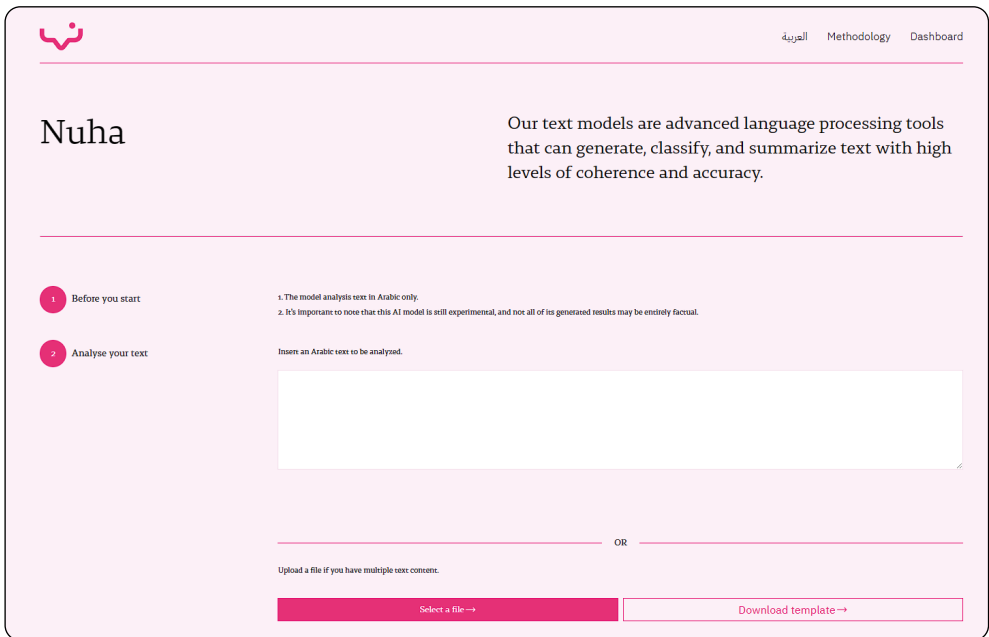


Figure 11: A screenshot of Nuha's Dashboard

The project is open-source and welcomes community contributions, especially those that go beyond its initial goals. This can strain resources, however. While openness is crucial, the demand for time, expertise, and infrastructure could pose challenges in the future. It is essential to balance open-source values with resource needs, so as to maintain the tool's value and sustainability for everyone involved.

Future Improvements

JOSA sees potential for improving the AI model Nuha, considering advancements in technology, including through:

- **Expanding the model classification capacity:** Currently, Nuha only identifies whether or not content is hate speech. Enhancing its ability to recognise various forms of hate speech, offensive language, and other non-hate speech categories would provide users with a better understanding of the content they are analysing.
- **Improving detection accuracy:** By consistently fine-tuning the model's algorithms and exposing it to larger, more diverse datasets, we can significantly enhance its ability to accurately identify instances of hate speech.
- **Intensity-scale classification:** The intensity-scale classification feature will provide users with information about the intensity of identified hate speech, based on Figure 02 – classifying and identifying the intensity of hate speech – and recommend actions to prevent potential harm.
- **User-generated feedback:** User-generated feedback allows users to provide feedback on the annotated data, so as to improve or enhance the results of the model. This user feedback helps make the annotation better, by incorporating a broader range of feedback and improving the model's accuracy.
- **Multi-dialect support:** Nuha was mainly trained on the Jordanian context and dialect. Expanding its ability to detect hate speech in other Arabic dialects from the MENA region, such as Levantine, Iraqi, and Egyptian, would significantly increase its reach and practicality.
- **Real-time monitoring:** Implementing real-time monitoring features would allow Nuha to quickly identify and respond to emerging hate speech trends.
- **Integration with social media platforms:** Connecting with different social media platforms, including Meta and X, to discover possibilities for integrating and adapting Nuha to interact to posts, comments, tweets, and retweets, will allow for real-time monitoring and enable Arabic-content moderators.
- **Context-aware classification:** Adding context-aware classification will provide additional information or context analysis

when identifying hate speech, through providing definitions of specific resources, words, similar examples, or additional context and categorisation.

- **Establish an OGBV research hub:** The current availability of research papers and educational materials related to OGBV topics in Arabic is limited. Establishing an OGBV research hub, through collaboration among TAMAM coalition members and partnerships with universities, research institutions, and NGOs, utilising Nuha, could significantly increase the quantity of educational and research papers in Arabic within Jordan.
- **Publish awareness materials:** Producing and publishing awareness material in the Arabic language that focus on project findings, including OGBV and AI, and OGBV and social media platforms topics, would make more information available on the subject.
- **OGBV Monitoring in Jordan:** This would involve positioning Nuha to become the main OGBV monitor tool used by research units to detect and report on ongoing incidents of OGBV.

Conclusion

Nuha is one of the tools JOSA has adopted in addressing the OGBV and promoting a safer digital space, as it identifies OGBV in Arabic, and

specifically the Jordanian dialect. In addition, the tool raises awareness about the prevalence and impact of OGBV, and promotes discussions and efforts to ensure a safer online space in Jordan.

The process extended beyond social media monitoring of 60 women's accounts, dataset preparation, and code development; it involved a collaborative effort that encompassed various quality checking techniques, testing, and exploration of alternative methods, requiring a dedicated team. Throughout the development of Nuha, DRI, TAMAM members, and JOSA played crucial roles in a collaborative journey. Through these, JOSA successfully developed a model with a 72 per cent (0.72/1) F1 score. The team remains dedicated to further enhancing the model, with the aim of raising the F1 score by increasing the size of the dataset.

Recommendations and Lessons Learned

Based on the insights and experiences from building an AI model to detect OGBV in the Jordanian dialect, we recommend the following for developers, teams, and stakeholders:

- **Digital archiving:** Researchers are advised to utilise digital tools like the WayBack Machine to archive content. Due to its limitations, however, we recommend the development of specialised digital archiving tools tailored for OGBV research.

- **Greater transparency by social media platforms:** There is a lack of clarity on social media platforms' content moderation processes, especially regarding OGBV. We urge these companies to provide in-depth insights into their content moderation methods, particularly for the Arabic content.
- **Data accessibility:** Platforms such as X should offer more public access to their data. We advocate for X to either make their APIs freely available for research purposes or to significantly reduce their pricing.
- **Industry-academia collaboration:** The gap between the tech community and academia is evident, especially in the areas of OGBV and the Arabic language. We recommend fostering closer ties between these two sectors to promote knowledge sharing and collaborative projects.
- **Partnership with rights organisations:** Collaborations with women's rights and OGBV-focused organisations are invaluable. We propose strengthening and sustaining these partnerships, emphasising the potential for broader pan-Arab cooperation between tech communities and civil society organisations.
- **Investment in AI and Arabic-language expertise:** There is a limited number of experts on the niche intersection of AI and the Arabic language, many of whom are in the private sector. We suggest funneling financial support to civil society initiatives to hire proficient AI specialists. Additionally, investments in education and training are vital to expanding the pool of qualified professionals in this area.
- **Data sample sizes:** To improve the effectiveness of AI in detecting hate speech against women, it is suggested to first acquire data on hate speech in the Arabic context, and then specialise it for targeted hate speech against women in the region.

References

1. **International Center for Research on Women**, (2018), TECHNOLOGY-FACILITATED GENDER-BASED VIOLENCE: WHAT IS IT, AND HOW DO WE MEASURE IT, https://www.svri.org/sites/default/files/attachments/2018-07-24/ICRW_TFGBVMarketing_Brief_v8-Web.pdf.
2. **United Nations**, (2023), Understanding Hate Speech, <https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech>.
3. **Paasch-Colberg, Sünje**, (2022), "Insults, Criminalisation, and Calls for Violence: Forms of Hate Speech and Offensive Language in German User Comments on Immigration.", https://doi.org/10.1007/978-3-030-92103-3_6.
4. **European Institute for Gender Equality**, (2023), EIGE's Gender Equality Glossary & Thesaurus is a specialised terminology tool focusing on the area of gender equality, https://eige.europa.eu/publications-resources/thesaurus/terms/1286?language_content_entity=en.
5. **IBM**, What is machine learning?, <https://www.ibm.com/topics/machine-learning>.
6. **Thomas Wood**, <https://deepai.org/machine-learning-glossary-and-terms/f-score>.
7. **The Jordanian National Commission for Women**, (2022), Violence Against Women in the Public and Political Spheres in Jordan, <https://link.josa.ngo/9Ythud>.
8. **Paasch-Colberg, Sünje**, (2022), "Insults, Criminalisation, and Calls for Violence: Forms of Hate Speech and Offensive Language in German User Comments on Immigration.", https://doi.org/10.1007/978-3-030-92103-3_6.
9. **Jordan Open Source Association**, (2022), Nuha's Annotation Guideline, <https://docs.google.com/document/d/1TPC0Fmg6UqXN7PVCXsXLyR85weyIi0bkb-KjrufAOv0/edit>

10. **Neag School of Education – University of Connecticut** (2021), Educational Research Basics, <https://researchbasics.education.uconn.edu/confidence-intervals-and-levels/#:~:text=The%20confidence%20level%20tells%20you%20how%20sure%20you,confidence%20level%20means%20you%20can%20be%2099%25%20certain.>
11. **Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova** (2019), Computation and Language (cs.CL), <https://arxiv.org/abs/1810.04805>
12. **Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz; Polosukhin, Illia** (2017), Attention is All you Need, [https://en.wikipedia.org/wiki/Transformer_\(machine_learning_model\)#cite_note-2017_Attention_Is_All_You_Need-1](https://en.wikipedia.org/wiki/Transformer_(machine_learning_model)#cite_note-2017_Attention_Is_All_You_Need-1)
13. **Datacamp** (2022), A Complete Guide to Data Augmentation, <https://www.datacamp.com/tutorial/complete-guide-data-augmentation>
14. **Hirsch, Eric Donald** (2002), The New Dictionary Of Cultural Literacy: What Every American Needs to Know (3rd ed.), https://en.wikipedia.org/wiki/Web_browser#cite_note-12-1
15. **Martin** (2011), API Design for C++ Reddy, API Design for C++, <https://books.google.com/books?id=IY29LyIT85wC>
16. **Open API documentation** is available at: Nuha API - Swagger UI.
17. **Detailed instructions for running the deployment environment** are uploaded to JOSA's GitHub repos, as referenced below:
 - a. Jordan Open Source Association (JOSA), (2023), NUHA's web client source code, <https://github.com/jordanopensource/nuha-web>.
 - b. Jordan Open Source Association (JOSA), (2023), NUHA's API client source code, <https://github.com/jordanopensource/nuha-api>.
 - c. Jordan Open Source Association (JOSA), (2023), NUHA's Huggingface Dataset model, <https://huggingface.co/thejosango/nuha/tree/main>.

Glossary

Introduction

The Jordan Open Source Association (JOSA) is currently developing an artificial intelligence (AI) model called Nuha, named after the Arabic word for "mind" or "brain." The model's main purpose is to assist researchers in identifying online gender-based violence against women in Jordan, particularly on social media platforms.

The purpose of this glossary document is to furnish definitions for all technical terminology employed during the tool's development process. It is important to highlight that the glossary will undergo periodic updates to encompass a wide range of aspects related to the model.

The glossary document encompasses three primary categories of definitions, as outlined by JOSA. These are:

1. Research-based terminology, which exclusively covers terminologies associated with the discovery phase of the tool, considering that research constituted a significant portion of this phase;
2. Technical-based terminology, which concentrates on all technical terms linked to the development of the model; and
3. General terminologies, which encompass terms encountered by JOSA during the regional reporting on social media and while working on the Nuha website.

General Terminologies

TERMS	DEFINITIONS
Honour Killing	A practice with a history in Jordan (and other countries in the region) that is entrenched in colonial legal systems that grant lenient sentences for the perpetrators, while using the victim's alleged "unethical behaviour" as justification. Since women would be vulnerable to physical harm without any legal (and, occasionally, societal) protection, women are reluctant to file formal complaints, even if this amounts to the danger of their being killed.
Violent Criticism	The act of denouncing; publicly menacing or making accusations; the act of inveighing against, stigmatising, or publicly arraigning someone.

Research-Based Terminologies

TERMS	DEFINITIONS
Crowdtangle	A tool from Meta that helps in following, analysing, and reporting on what is happening across social media, by tracking engagement (interactions), which includes reactions (e.g., “Likes” in Facebook), comments, and shares. ¹⁶
Export Comment	A tool that is used to scrape content from social media platforms, such as Facebook or Twitter. ¹⁷
Hate Speech	Offensive discourse targeting a group or an individual based on inherent characteristics (such as race, religion or gender) and that may threaten social peace. ¹⁸
Offensive Language	Language that does not qualify as hate speech, but still escalates online discussions about gender and women’s rights. ¹⁹
Online Gender-Based Violence (OGBV)	An action by one or more people online that harms others based on their sexual or gender identity, or by enforcing harmful gender norms. ²⁰
Social Media	A digital technology that allows the sharing of ideas and information, including text and visuals, among virtual networks and communities. ²¹
Twitter API	A set of programmatic endpoints that can be used to understand or build the conversation on Twitter. The API (application program interface) allows us to find and retrieve, engage with, or create a variety of different resources, including tweets, users, and spaces.
Woman	A person assigned female sex at birth, or a person who defines herself as a woman. ²²

¹⁶ <https://www.crowdtangle.com/>

¹⁷ <https://exportcomments.com/>

¹⁸ Understanding Hate Speech”, United Nations. <https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech> (Accessed May 2023 ,28)

¹⁹ Paasch-Colberg, Sünje, et al. “Insults, Criminalisation, and Calls for Violence: Forms of Hate Speech and Offensive Language in German User Comments on Immigration.” Cyberhate in the Context of Migrations, 2022, pp. 63–137, https://doi.org/10.1007/978-3-92103-030-6_3-92103-030. Accessed 20 Nov. 2022.

²⁰ International Center of Research on Women https://www.svri.org/sites/default/files/attachments/24-07-2018/ICRW_TFGBVMMarketing_Brief_v-8Web.pdf

²¹ <https://www.investopedia.com/terms/s/social-media.asp> (Investopedia, 2023)

²² https://eige.europa.eu/publications-resources/thesaurus/terms/1286?language_content_entity=en

Technical-based Terminologies

TERMS	DEFINITIONS
Application Program Interface (API)	A way for two or more computer programs to communicate with each other. A type of software interface, offering a service to other pieces of software. ²³
Artificial Intelligence (AI)	Computer systems that are able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages. ²⁴
BERT	Bidirectional Encoder Representations from Transformers. ²⁵ It is a sophisticated language model in the field of artificial intelligence that can understand the meaning of words in a sentence by analyzing the words that come before and after them, enabling it to grasp the context and nuances of language more effectively.
Confidence Level	A measure that tells us how sure we can be about the accuracy of the annotations made by a computer model. It is like a measure of how much trust we can have in a model's decisions. The higher the confidence level, the more certain we are that the model's annotations are correct. ²⁶
Data Augmentation	A technique of artificially increasing a training set by creating modified copies of a dataset using existing data. It includes making minor changes to the dataset or using deep learning to generate new data points, i.e., a number of techniques used to generate new datasets based on original data. ²⁷
F1-score	A metric used to evaluate a machine-learning model by measuring its accuracy. ²⁸

²³ <https://books.google.com/books?id=IY29LyIT85wC>, API Design for C++ Reddy, Martin (2011).

²⁴ <https://www.sciencedirect.com/topics/social-sciences/artificial-intelligence> (Science Direct, 2018)

²⁵ https://d2l.ai/chapter_attention-mechanisms-and-transformers/large-pretraining-transformers.html (Dive into Deep Learning, 2023)

²⁶ <https://researchbasics.education.uconn.edu/confidence-intervals-and-levels/#:-:text=The20%confidence20%level20%tells20%you20%how20%sure20%you,confidence20%level20%means20%you20%can20%be20%25%2099%certain>. (University of Connecticut (2021), Educational Research Basics)

²⁷ <https://www.datacamp.com/tutorial/complete-guide-data-augmentation> (Datacamp, 2022)

²⁸ <https://deeplai.org/machine-learning-glossary-and-terms/f-score>

TERMS	DEFINITIONS
False Negative (FN)	An outcome where the model incorrectly predicts the negative class ²⁹ (i.e., when the model fails to recognise that content contains hate speech and labels it as hate speech-free).
False Positive (FP)	An outcome where the model incorrectly predicts the positive class ³⁰ (i.e., when the model labels hate speech-free content as hate speech).
Fine-tuning	A common technique for transfer learning. The target model copies all model designs, except the output layer, with their parameters from the source model, and fine-tunes these parameters based on the target dataset. In contrast, the output layer of the target model needs to be trained from scratch. ³¹
Hugging Face	A machine learning and data science platform and community that helps users build, deploy, and train machine-learning models. ³²
Machine Learning	A branch of artificial intelligence (AI) that focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. ³³
Open Source Software	Software with source code that anyone can inspect, modify, and enhance. ³⁴
Transformer-based model	A deep learning architecture, initially proposed in 2017, that relies on the parallel multi-head attention mechanism, which allows for processing and analysing different aspects of input data (language) simultaneously, which increases efficiency. ³⁵
Under-sampling	Reducing the amount of overrepresented data to make it more equal with the underrepresented data.
Web Client (browser)	A web client (more commonly referred to as a browser) is an application for accessing websites and the internet. ³⁶

²⁹ [Google Developers' Site](#).

³⁰ [Google Developers' Site](#).

³¹ https://d2l.ai/chapter_computer-vision/fine-tuning.html (Dive into Deep Learning, 2023)

³² <https://huggingface.co/>

³³ <https://www.ibm.com/topics/machine-learning>

³⁴ <https://opensource.com/resources/what-open-source>

³⁵ [https://en.wikipedia.org/wiki/Transformer_\(machine_learning_model\)](https://en.wikipedia.org/wiki/Transformer_(machine_learning_model))

³⁶ https://en.wikipedia.org/wiki/Web_browser#cite_note-1-12 The New Dictionary Of Cultural Literacy: What Every American Needs to Know (3rd ed.) Hirsch, Eric Donald (2002).

Fourth Report

Words Matter Network Regional Forum Report: Expert Insights on MENA's Digital Landscape

Executive Summary

DRI's "Words Matter Network Regional Forum" on disinformation and hate speech in the MENA Region gathered leading experts, organisations, and stakeholders to delve into crucial challenges and opportunities shaping the digital landscape. Diverse sessions examined multifaceted issues, such as countering online violence, promoting freedom of expression, addressing hate speech, and analysing the information environment in the region.

Throughout the event, numerous recommendations emerged, advocating for long-term legislation to counter information disorder,

and stressing the need for fact-checking mechanisms, media literacy promotion, and enhanced collaboration among stakeholders. Participants highlighted the misuse of penal codes against journalists and activists, emphasising the importance of protecting freedom of speech while navigating legal challenges. Common themes across sessions emphasised the role of social media platforms in addressing online harm and ensuring accountability. Challenges related to the regulation of social media algorithms, understanding cultural contexts, and promoting inclusivity in online spaces were raised regularly during the discussions. It was evident during these discussions that challenges in managing online

violence, disinformation, and hate speech persist. Opportunities for improvement identified were centred on fostering collaboration among stakeholders, enhancing media literacy, and creating more inclusive digital environments.

The forum concluded with a call for continued cooperation, the adoption of robust legislation, and the promotion of media literacy to address challenges and seize opportunities in shaping a more equitable and responsible digital landscape across the MENA region.

Introduction

The "Words Matter Network Regional Forum" was held virtually from October 31st to November 2nd, 2023, with a focus on disinformation, hate speech, and online harm in the MENA region. The event convened a diverse group of thought leaders, journalists, activists, academics, and policymakers. The forum sought to present solutions, foster dialogue, and achieve resilience in the face of increasing challenges posed by harmful behaviors in the digital sphere.

Initial plans for an in-person gathering in Beirut were disrupted by the Israel-Hamas war and the ongoing violent escalation between Hezbollah and Israel in the south of Lebanon, and the event was ultimately held virtually, bringing together around 200 people from different countries and regions around the globe. Over the course of 11 sessions, the forum

included comprehensive discussions, workshops, and expert dialogues.

Drawing upon the expertise of over 20 speakers from Words Matter partner organisations, eight external entities, and three independent experts, and by exploring different themes, such as online hate speech, cyberbullying, and the spread of disinformation during key political moments on social media, the event produced many rich insights and approaches.

This summary report captures the essence of these dialogues, including key insights, recommendations, and main trends and conversations discussed during the event. It serves as a compendium of collective wisdom, fostering continued momentum in the fight against online misinformation, hate speech, and digital malpractices. Through this summary, we aim to promote knowledge-sharing, raise awareness, identify emerging trends, empower resilience against hate speech and disinformation on social media, inspire action, and encourage collaboration among stakeholders.

Day One

Session 1

AI Solutions Detecting and Countering Online Violence and Manipulation in the MENA Region

Day Two

Session 1

Countering Online Violence in the MENA Region: Case Studies

Day Three

Session 1

Informational Disorder and Political Transitions: Lessons Learned from the Arab World and Post-25th July 2021 in Tunisia

Session 2

How to Expose a Network: Identifying and Analysing a Disinformation Campaign

Session 3

The Technical Challenges in the Usage of AI Tools within the New API and META Policies

Session 4

Understanding the Landscape of the Information Environment in the WANA Region: Trends and Challenges

Session 2

Promoting Freedom of Expression while Countering Hate Speech, Disinformation, and Gender-Based Violence in the MENA Region: The Importance of Multi-Stakeholder and Multi-Faceted Approaches

Session 3

The EU Digital Services Act and Opportunities for Adaptation in the MENA Region: The League of Arab States Social Media Strategy as an Example

Session 2

Social Media Platforms in the MENA Region: Addressing Online Harms and Accountability

Session 3

Information Environment in Lebanon and Iraq

Session 4

The Misuse and Malpractices of Penal Codes Towards Journalists, and Activists' Oppression in The Digital Space during Elections

Day One Sessions

Day Two

Day Three

Day 1, Session 1 – AI Solutions Detecting and Countering Online Violence and Manipulation in the MENA Region

In the opening session of the forum, addressing hate speech and disinformation in the MENA region, speakers highlighted collaboration among countries and organisations, acknowledging the challenges posed by regional instability. The discussions centred on AI's role in monitoring and countering hate speech and violence on social media platforms. A comparison presented by 7amleh on the different uses of their AI powered platform (7or) between Hebrew and Arabic showed that Arabic is more conducive to detecting hate speech and violence, with examples presented indicating that Arabic lends itself more readily to such analysis when compared to Hebrew. Presenters also showcased tools like Nuha, developed by JOSA, a Words Matter Network partner, designed to detect hate speech targeting women, and eMonitor+'s suite of AI tools for media platform monitoring, developed by the UNDP, that can help journalists and researchers build automated verification and fact-checking workflows powered by AI to counter disinformation on a regional level in the MENA region. They highlighted challenges, such as data scarcity and the need for human intervention, in AI-driven analysis. Questions centred on maintaining freedom of speech boundaries while utilising AI and ensuring data accuracy and classification, suggesting a need for the ICT community to create more cost-effective AI tools. This session emphasised

leveraging AI to tackle online violence and hate speech, showcasing initiatives like Nuha and eMonitor+'s tools. It also underscored the importance of ensuring accuracy in AI-driven analyses and the need for collaboration to effectively address challenges in the MENA region.

Day 1, Session 2 – How to Expose a Network: Identifying and Analysing a Disinformation Campaign

The session focused on identifying and analysing disinformation campaigns, with discussions emphasising coordinated efforts to spread false information versus authentic user campaigns. Examples showcased manipulation through fake accounts, changing narratives, and the impact on conversations. Presentations from various projects in the MENA region highlighted techniques used by coordinated groups, involving hidden networks and real-time content presentation, emphasising the need to deconstruct disinformation attempts.

Day 1, Session 3 – The Technical Challenges in the Usage of AI Tools within the New API and META Policies

The session was centred on the challenges related to AI tools and new policies. The discussions involved the impact of AI-generated disinformation on electoral processes in Turkey and Slovakia. Tools like Maharat Tracking were used for monitoring legislative elections in Lebanon, albeit with some disappointment in the results. Challenges included the need for identity protection tools during the analysis of

harmful content, considerations for illiterate users and Arabic dialects, and ensuring the reliability of data in the MENA region. The session also reflected on how Generative AI is changing the current disinformation campaigns, by adding synthetic media created by AI that makes it harder to detect and verify by researchers. During this session, DRI presented the emerging research of the Digital Democracy team.

Day 1, Session 4 – Understanding the Landscape of the Information Environment in the WANA Region: Trends and Challenges

The session delved into the information landscape in the WANA region, highlighting challenges faced in monitoring, gender discourse, and reports for public awareness. Discussions emphasised freedom of speech, media roles, and trust in traditional media in Tunisia. Key points were the impact of social media in the MENA region, disinformation during elections and the COVID-19 pandemic, and the Arab Spring's influence. The questions and answers portion of the session covered challenges in creating a strong media environment, judicial responses to urgent hate speech cases, and complexities in analysing hate speech content. Suggestions included the need for stronger laws against hate speech, further tool development, networking for leveraging scientific papers, and the role of investigative journalism in countering disinformation.

Day One

Day Two Sessions

Day Three

Day 2, Session 1 – Countering Online Violence in the MENA Region: Case Studies

The session commenced with an introduction on cyber violence in the MENA region, stressing its continuity into real-life violence. Panelists highlighted attacks on women in public spaces, and on activists, journalists, and politicians on social media, and emphasised the need to protect these individuals. Recommendations included efforts by social media companies, civil society, and religious institutions to combat online violence. The discussion covered hate speech, gender-based violence, and the protection of researchers involved in this work. Panelists presented case studies on violence against activists, politicians, and migrants. The analysis categorised hate speech into stereotypes, dehumanisation, and expressions of violence. Recommendations urged collaboration among stakeholders in Jordan and actions to counter hate speech targeting African migrants in Tunisia. The session concluded with calls on technical researchers to collaborate in creating glossaries of violent language and safeguards against fake profile use for online defamation. Discussions highlighted the pervasive attacks against women across social media platforms.

Day 2, Session 2 – Promoting Freedom of Expression while Countering Hate Speech, Disinformation, and Gender-Based Violence in the MENA Region: The Importance of Multi-Stakeholder and Multi-Faceted Approaches

The session focused on promoting freedom of expression while countering hate speech, disinformation, and gender-based violence in the MENA region through multi-stakeholder approaches. Presenters highlighted the need for defining hate speech and legal frameworks while advocating participatory approaches. Recommendations included emphasising gender equality on social media, fostering dialogue among stakeholders, and training analysts in AI techniques. Different speakers stressed the importance of understanding international agreements and the impact of political agendas on media content. They recommended early education on social media literacy, and involving religious figures to address misuse of online data. The session also tackled the complexities of defining hate speech and the need for software adapted to the Arabic context. In the question-and-answer portion of the session, speakers discussed defining hate speech, advocating for minority rights, and involving local experts in technology solutions. They emphasised the need for a common definition of hate speech in the Arab world, and for proactive involvement in combating violence online.

Day 2, Session 3 – The EU Digital Services Act and Opportunities for Adaptation in the MENA Region: The League of Arab States Social Media Strategy as an Example

The session addressed the European Union's Digital Services Act's impact on the MENA region and the possibilities for its adaptation. Panelists examined how the regulatory landscape influences social media platforms, emphasising freedom, transparency, and challenges in implementing the DSA outside the EU and the United Kingdom. Examples were shared about governments regulating internet usage, citing concerns like false news, electoral interference, and incitement to unrest. The discussion highlighted potential risks in the MENA region, such as censorship and the imprisonment of activists under the pretext of DSA implementation. Connections were drawn between European law and regional strategies, acknowledging governments' hesitation to relinquish control to external bodies and favoring swift action over private entity reliance. The need for clear global standards applicable to local contexts was emphasised, urging collaborative efforts among researchers to effectively navigate these concepts. The conversation focused on adopting participatory decision-making, considering the replication of European strategies, and ensuring balanced content removal without imposing undue censorship. The session concluded with a call for clear guidelines, arbitration, and caution against hasty censorship during data removal decisions.

Day One

Day Two

Day Three Sessions

Day 3, Session 1 – Informational Disorder and Political Transitions: Lessons Learned from the Arab World and Post-25th July 2021 in Tunisia

In a session discussing informational disorder and political transitions, speakers highlighted disinformation's impact on Tunisia's transitional phase and the media's role in shaping public opinion after a significant event like the 25 July coup. The absence of trustworthy information, disinformation's proliferation since the coup, and challenges faced by fact-checkers were key points raised. Recommendations included enhancing fact-checking, revisiting media strategies, and being cautious of the disinformation agenda. Participants stressed the need for an approach to manage disinformation, emphasising the importance of updated, reliable information, and the role of civil society in correcting media disinformation. Suggestions were made to improve data verification, engage reliable sources, and conduct on-site training sessions. Additionally, they highlighted the divide between elite and grassroots groups, emphasising the need for more selective data issuance and careful engagement with online sources for greater impact.

Day 3, Session 2 – Social Media Platforms in the MENA Region: Addressing Online Harms and Accountability

In a session focusing on social media platforms in the MENA region, speakers highlighted the need for accountability and transparency. They addressed concerns about biased algorithms, the responsibility of ICT companies, and the impact of social media on various regions. Discussions emphasised accountability for social media platforms, distinguishing between data consumers and producers, and investing in AI to bridge gaps in social media outputs. Concerns were raised about social media's impact on mental health, the financial harm it causes, and the need to combat hate speech. Speakers recommended clarifying data processes, enhancing collaboration to combat hate speech, understanding the role of algorithms in controlling narratives, and acknowledging shared responsibility between consumers and content producers. Questions arose about fair practices by ICT companies and strategies to counteract unfairness in their practices. Suggestions were made to ensure transparency in platform dealings, to understand cultural values in software development, and to define concepts like hate speech and violence more clearly. The session concluded with a reminder of interconnected concepts like accountability, monitoring, censorship, and funding, underscoring the need to address these issues in tandem within the social media landscape.

Day 3, Session 3 – Information Environment in Lebanon and Iraq

In a session dedicated to understanding the information environment in Lebanon and Iraq, panelists highlighted the prevalent challenges of disinformation and hate speech in both countries. They pointed out that misinformation often originates from social media sources backed by neighboring countries, and how this significantly influences public perception on various issues. Recommendations centred on establishing long-term legislative countermeasures, promoting media independence, and enhancing fact-checking protocols. The discussions focused on the diverse media landscape in Lebanon, where misinformation through online platforms plays a pivotal role in shaping public opinion. In Iraq, challenges related to government data access, digital transition, and electoral misinformation were highlighted, especially in relation to the country's multicultural society. The session also addressed the financial constraints impacting Iraq's media sector, where many outlets are funded by political entities, influencing discourse on social media platforms during elections. The question-and-answer portion of the session touched upon the selection of countries for study and the emphasis on electoral processes in related work, primarily aimed at empowering youth in election-related capacities.

Day One

Day Two

Day Three Sessions

Day 3, Session 4 – The Misuse and Malpractices of Penal Codes Towards Journalists, and Activists’ Oppression in The Digital Space during Elections

During this session, panelists discussed the misuse of penal codes to suppress journalists and activists, particularly during elections. Concerns were raised about the misuse of laws to curb freedom of expression and target social media users. Cases from various countries, such as Egypt and Saudi Arabia, highlighted smear campaigns and legal actions against activists advocating for free speech. Presenters emphasised the need for protective measures for journalists and activists, citing cases of their imprisonment and legal harassment against them for expressing opinions or criticising authorities. They underscored the importance of distinguishing between genuine charges and freedom of speech violations, and urged clarity regarding laws that suppress journalists' rights. Recommendations included the need to review and revise penal codes threatening journalists, emphasising the role of trade unions in protecting journalists' rights, and calling for transparency in media practices. Concerns about specific legal articles targeting journalists were addressed, advocating for a re-examination of laws that pose threats to freedom of speech. The question-and-answer part of the session highlighted the

need for regulatory frameworks that support journalists' safety and freedom of expression. Panelists stressed the importance of expanding case studies beyond specific countries to understand wider threats to free speech. Overall, the session underscored the dangers posed by the misuse of legal measures against journalists and activists, and emphasised the need for protective measures to safeguard their rights and safety.

Cross-Session Themes

Across multiple sessions, several recurring themes and trends emerged. These recurring themes underscored the multidimensional nature of the challenges faced in the digital space, and emphasised the need for comprehensive, collaborative, and multi-stakeholder approaches to address them effectively. The main thematic areas that emerged during the discussions were grouped under six different categories, as outlined below.

1. Legislation and Policy

Enhancement: Many sessions highlighted the need for long-term legislation and policy frameworks to counter disinformation, hate speech, and online violence. These suggestions emphasised international standards, fact-checking protocols, and legal reforms to protect freedom of expression while curbing the misuse of penal codes.

2. Media Integrity and

Independence: Discussions consistently stressed the importance of an independent media environment, supported by fact-checking mechanisms and media literacy initiatives. Recommendations aimed to safeguard media integrity, to encourage transparency, and to combat corrupt practices within the media landscape.

3. **Civil Society Engagement:** Suggestions often revolved around involving civil society organisations, NGOs, and various stakeholders in the process of addressing online harm. The sessions highlighted the need for collaborative efforts involving these groups to counter hate speech, promote accountability, and protect journalists and activists.
4. **Awareness and Education:** There was a strong emphasis on raising awareness and enhancing education regarding online risks, disinformation, and hate speech. Suggestions included promoting media literacy among users, providing training sessions for various stakeholders, and engaging religious leaders to address the misuse of online platforms.
5. **Transparency and Accountability of Tech Companies:** There was a consistent call for greater transparency and accountability from social media and tech companies. The speakers urged these companies to clarify their content moderation policies, involve civil society in decision-making processes, and adapt platforms to suit regional contexts and marginalised groups.
6. **Regional Collaboration:** Multiple sessions stressed the importance of regional cooperation and collaboration. Recommendations often highlighted the need for

joint efforts between countries in the MENA region to address online harm, share best practices, and develop common strategies to tackle disinformation and hate speech.

Recommendations

The event generated a range of recommendations to tackle the challenges of misinformation, hate speech, and online violence in the MENA region. These recommendations underscore the need for a comprehensive approach involving legal, educational, technological, and collaborative measures to address the complexities of online harm in the MENA region. The recommendations are grouped under eight thematic areas, as outlined below.

1. **Legislation and Policies:** Craft long-term legislation to counter information disorder, promoting policies against misleading narratives and establishing fact-checking protocols.
2. **Collaboration and Awareness:** Increase collaboration among stakeholders, including legislators, civil society, media outlets, and tech companies, to promote media independence and combat online harm.
3. **Media Literacy and Fact-Checking:** Promote media literacy among users, installing robust fact-checking mechanisms across media platforms and emphasising accuracy and reliability in

disseminating information.

4. **Freedom of Expression:** Safeguard freedom of expression, while distinguishing between actual attacks using hate speech and practices of disinformation and freedom of speech, and ensuring transparent and accountable media practices.
5. **Tech Engagement and Transparency:** Encourage tech companies to be transparent about their platform operations, to involve civil society organisations, and to address biases in algorithms to prevent online harm.
6. **Education and Training:** Conduct regular training sessions, on-site training, and awareness programmes to enhance skills in handling online information and addressing hate speech.
7. **International Collaboration:** Align with global efforts, revise strategies of international media organisations, and collaborate with experts from other regions to combat misinformation.
8. **Social Accountability:** Focus on social accountability more than individual accountability, promote collaboration between governments and citizens, and restore trust among stakeholders.

Conclusion

The event resulted in many conclusions that are important for understanding the digital landscape in the MENA region. Among these conclusions, the

prevalence of disinformation, hate speech, and misleading narratives emerged as pressing concerns, affecting the political, media, and societal fabric. It became evident that each country faces unique challenges, requiring tailored solutions within diverse cultural and political contexts. Collaborative efforts, encompassing governments, civil society, media, tech entities, and citizens, were underscored as essential for effectively combatting online harm.

Understanding the complexities of AI, navigating legal frameworks, and bridging the implementation gap emerged as crucial points of focus. Emphasising media literacy, fortifying fact-checking mechanisms, and upholding freedom of expression stood out as fundamental pillars in this pursuit. Balancing the fight against hate speech and misinformation while preserving freedom of expression is a delicate, yet vital consideration. Furthermore, technical enhancements, including transparent algorithms and enhanced tech accountability, were deemed necessary. Ultimately, a comprehensive, multi-stakeholder approach, coupled with ongoing efforts to implement actionable strategies, remains the path forward in addressing these challenges.

Acknowledgments

We extend our deepest gratitude to all of the contributors and participants whose involvement made this three-day forum a success. Special thanks go to Words Matter partner organisations

– the Al-Hayat Center (RASED), the Institute of Press and Information Sciences (IPSI), the Jordan Open Source Association (JOSA), the Maharat Foundation, and Mourakiboun – for their invaluable collaboration and commitment to this event.

We also express our sincere appreciation to the partnering institutions, including 7amleh, Article-19, DFR Lab, INSM-Iraq, NDI, SMEX, Tech4Peace, TuniFact, UNDP, and all of the esteemed speakers, panelists, and attendees who shared their expertise, insights, and experiences.

A warm acknowledgment goes to the event staff, including the moderator, the interpreters, and the sign language interpreter, whose dedicated support ensured smooth communication throughout the forum. Our gratitude extends to all project members and to our esteemed former colleagues Lena-Maria Böswald, Makrem Dhifali, and Amira Kridagh for their invaluable contributions.

We are grateful for the exploration of diverse sub-topics during the sessions, all of which enriched the depth and breadth of discussions and insights shared at this Forum. Thank you all for your invaluable contributions, dedication, and enthusiasm.

About Words Matter

Contact: menahub@democracy-reporting.org

DRI has been increasingly active in the field of social media monitoring (SMM) since 2017, strengthening local capacities to monitor social media during elections, sharing information and evidence gathered in different countries, bringing together expert organisations, producing methodologies, and informing public and expert debate.

Within the framework of the “Words Matter” project, DRI and its partners seek to contribute to strengthening the safeguarding of democratic processes and societies’ resilience to online disinformation and hate speech in the MENA region.

DRI works with partner organisations from four countries (Jordan, Lebanon, Sudan, and Tunisia), strengthening local capacities to monitor and analyse

online disinformation and hate speech during key national democratic processes, while building a regional network to allow for comparative analysis and peer learning.

“Words Matter” aims to achieve the following objectives:

- Capacity-building for project partners **to acquire institutional skills to design sound social media monitoring methodologies**, to effectively monitor disinformation and hate speech online, and to enhance evidence of the impacts of disinformation and hate speech online on civic or political participation and human rights.
- **Enhanced multi-stakeholder and regional engagement** to advocate against and combat online disinformation and

hate speech, through a civil society network, as well as through continuous exchanges on transparent regulations.; and

- In the countries of project partners, **improved awareness and resilience of civic target groups**, and concrete action by decision-makers to transparently combat online hate speech and disinformation.

About the Digital Democracy Program

Contact: info@democracy-reporting.org

DRI's Digital Democracy (DD) programme protects online democratic discourse by exposing disinformation, manipulation and hate speech, strengthening the capacity of CSOs for monitoring and advocacy, and ensuring appropriate and evidence-based responses from governments and tech companies.

DRI is well-positioned to address online threats and disinformation, due to its research on manipulated media content, deepfakes as potential disinformation tools, and its current focus on identifying new potential threats and emerging technologies in this field. As part of our diverse toolbox, we have, for example, integrated machine learning models to help us identify emerging trends in the disinformation space. Our work on information manipulation is also complemented by analysing and publishing guides on gender-based under-representation and harassment online.

An important activity within the DD programme for exposing and fighting hate speech and disinformation is social media monitoring (SMM). SMM is the objective analysis of democratic discourse and political actors on social media platforms. This is far more complex than traditional media monitoring, with a myriad of actors and content, combining official democratic institutions (e.g., political parties, politicians, media) and unofficial actors (e.g., individuals, political influencers, partisan groups). This is why DRI published the Digital Democracy Monitor Toolkit, the first social media monitoring methodology that helps civil society, journalists, and academia to research social media and democracy.

Our methodology was tested and used for conducting social media monitoring in 12 countries (including [Germany](#), [Libya](#), [Myanmar](#), Nigeria and [Sri Lanka](#)), focusing on disinformation, hate speech and political advertising before, during and after the elections. By using a holistic

approach to analyse social media, our toolkit engages with disinformation and hate speech by looking at the message or content, the active messengers, and the messaging, thus both the forms and the channels of distribution.

Based on the findings of our SMM, we have advocated for the implementation of the European Democracy Action Plan (EDAP) commitments, which could strengthen the fight against disinformation at the EU level and contribute to the debate about content-ranking systems, a major challenge when it comes to the dissemination of dis/misinformation. DRI has also lobbied for the implementation of the EU's Digital Service Act, a potential milestone in the effort to increase accountability across social media platforms. In launching the Arabic version of the SMM toolkit, we hope to empower the MENA region in the same way.

About DRI

Contact: info@democracy-reporting.org

Democracy Reporting International (DRI) is an independent organisation dedicated to promoting democracy worldwide. We believe that people are active participants in public life, not subjects of their governments. We strengthen democracy by supporting the institutions and processes that make it sustainable, and work with all stakeholders towards ensuring that citizens play a role in shaping their country. Our vision is grounded in globally agreed upon principles of democracy, stemming from the democratic governance championed by the United Nations and international law.

DRI's work focuses on five key themes of democracy: Justice, Elections, Local Governance, Digital Democracy and Human Rights. By working at both the national and local level, we use five intervention approaches in our projects: awareness-raising, capacity-building, fostering engagement between different stakeholders, supporting the building of democratic institutions, and advising on the drafting and implementing of policies and laws.

DRI's work is led by a Berlin-based executive team and supervised by an independent board of proven

democracy champions. DRI maintains country offices in Lebanon, Libya, Tunisia, Pakistan, Myanmar, Sri Lanka and Ukraine. Through our networks of country offices and partners, we are in a unique position to track, document, and report developments and help make tangible improvements on the ground.

About DRI Partners

Al-Hayat Center for Civil Society Development: is a non-governmental civil society organization founded in 2006. The center has expanded to become one of the leading NGOs in Jordan. Al-Hayat's overall mission is to promote accountability, governance, public participation, and tolerance in Jordan and the region within the framework of democracy, human rights, the rule of law, and gender mainstreaming in public policy and actions.

Jordan Open Source Association (JOSA): is a non-profit organization based in Amman, Jordan. The association is among the few non-profits registered under the Jordan Ministry of Digital Economy and Entrepreneurship. JOSA's mission is to promote openness in technology and to defend the rights of technology users in Jordan. JOSA believes that information that is non-personal – whether it's software code, hardware design blueprints, data, network protocols and architecture, content – should be free for everyone to view, use, share, and modify. JOSA's belief also holds that information that is personal should be protected within legal and technological frameworks. Access to the modern Web should likewise remain open.

Lab'TRACK: is a laboratory for monitoring, analysis and reflection on political disinformation phenomenon on social networks, in particular the Facebook network. The laboratory is a collaboration between Mourakiboun and IPSI.

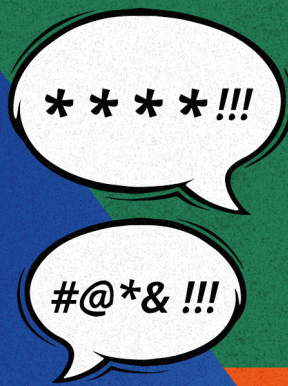
MOURAKIBOUN: Mourakiboun is a domestic electoral observation network that was launched in 2011 and is today a key player in this field with multiple national and international partners. Since 2014, Mourakiboun has been diversifying its actions by adding accountability of public services and support to the Tunisian decentralization process to its portfolio. Mourakiboun has a network of over 100 volunteers in all regions of Tunisia and excellent access to local structures and stakeholders. Mourakiboun has adopted an IT approach to its activities, thereby increasingly reaching Tunisian youth. During the 2014 and 2019 presidential elections, Mourakiboun conducted social media monitoring activities focused on the interactions of FB users with the speeches of candidates during electoral campaigns.

Institut de Presse et des Sciences de l'Information (IPSI): was established in 1967 and became a non-departmental public institution enjoying financial autonomy and legal personality in 1973. The Institute is known as Tunisia's leading university for the education of journalists and media workers. IPSI's research in the field of information and communication sciences has been met with international acclaim. IPSI has a network of national (INLUCC, HAICA, UFP) and international partners (Deutsche Welle Akademie, UNESCO, UNDP, Article 19 among others). Through this cooperation, IPSI provides specialized training sessions and hosts experts and internationally renowned speakers to introduce students to innovative practices in the field of communication.

MAHARAT: a women-led, Beirut-based organization, working as a catalyst, defending and advancing the development of democratic societies governed by the values of freedom of expression and respect for human rights.

Maharat advances the societal and political conditions that enhance freedom of expression and access to information, both online and offline. Maharat engages and equips a progressive community in Lebanon and the MENA region with the skills and knowledge necessary to create change.

Sudanese Development Initiative (SUDIA): In light of the ongoing armed conflict in Sudan, our valued partner from Sudan, The Sudanese Development Initiative (SUDIA), has regrettably been unable to continue their participation in the project. We deeply appreciate the significant contributions they made during their active involvement. While they are no longer with us due to these challenging circumstances, their dedication and expertise have left a lasting impact on the project's progress. As we move forward, we honor their commitment and extend our hopes for a peaceful resolution to the conflict in Sudan.



DEMOCRACY
REPORTING
INTERNATIONAL

Democracy Reporting International (DRI) was founded in 2006 by an international group of experts on democratic governance and elections.

DRI works on research and analysis to direct engagement with partners on the ground to improve democratic structures and safeguards across the countries where we work.

Elbestraße 28/29 12045 Berlin, Germany
info@democracy-reporting.org
wordsmatter@democracy-reporting.org
www.democracy-reporting.org/